THE BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ALAN STUART
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
E. A. PEEL
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
W. STEPHENSON
P. E. VERNON

Managing Sub-editor J. W. WHITFIELD

BIBL.
LONDIN.
UNIV.

VOLUME X

Printed and Published by
TAYLOR & FRANCIS LTD.
RED LION COURT, FLEET STREET, LONDON, E.C.4

Digitized by the Internet Archive
in 2026

https://archive.org/details/british-journal-of-mathematical-statistical-psychology_1957-05_10_part-1

Vol. X Part I The British Journal of Statistical Psychology May 1957

A STUDY OF THE PERFORMANCE OF 2,000 CHILDREN ON FOUR VOCABULARY TESTS

II. Norms, with some Observations on the Relative Variability of Boys and Girls

By M. I. Dunsdon and J. A. Fraser Roberts
Burden Mental Research Department, Stoke Park Hospital, Bristol

I. *Introduction.* II. *Increment of Mean Score with Age.* III. *Increment of Variance of Score with Age.* IV. *Norms.* V. *Efficiency of Adjusted Scores.* VI. *Frequency Distribution of Vocabulary Score.* VII. *Further Examination of Variability of Score in Boys and Girls.* VIII. *Summary.*

I. Introduction

Reasons have been given (Dunsdon and Roberts [1]) for choosing vocabulary scales as quickly-given individual tests, for the purpose, not of assessing single children, but of comparing large groups. For the comparison of groups vocabulary tests have manifest advantages in respect of reliability, validity, relative absence of practice effect, and above all, the amount of information obtained per unit of testing time. The four oral definition vocabularies used here are those included in the Terman–Merrill (1937 Revision) [8] Form L, and Wechsler's Intelligence Scale for Children [9], together with Raven's Mill Hill Vocabulary Tests A and B [4]. In the form for oral definition the Mill Hill vocabularies were designed by Raven to be used in combination. Accordingly in this paper results and norms are given in terms of the combined scores for M.H.A. and M.H.B. Norms are therefore presented, by sexes, for

1. Terman–Merrill Intelligence Scale, Form L (T.M.L.);
2. Wechsler Intelligence Scale for Children (W.I.S.C.);
3. Mill Hill A and Mill Hill B Vocabulary Scales combined (M.H.A. and M.H.B.);
4. All four vocabularies combined.

The reason for calculating norms is to derive scores which are independent of age and sex. The requirements for assessing individuals on the one hand and comparing large groups on the other are very different. For the latter a much higher standard of accuracy is needed. Relatively small systematic sources of bias may matter little in an individual measurement, but are important when the problem is (to quote a concrete example) whether twins are or are not slightly lower in average performance than their sibs, who are necessarily of different age. The point is well illustrated in a recent valuable review article by Yates [10], dealing with the use of vocabulary in the measurement of intellectual deterioration.

He mentions the higher average score of boys, quoting our study and others; and then adds: "The differences were so slight as to be unimportant". Actually, on taking the four vocabularies combined, at 6 years boys have a mean score of 26 words and girls 23; at 14 years boys score 76 and girls 71. In terms of 'mental ages' these are differences of about 5 months in young children and nearly a year in children approaching 15. Such differences may be unimportant in relation to the special problem Yates discusses; but they are of the order of magnitude of some of the differences between large groups of subjects that we had in mind when undertaking this work. The same considerations would apply to individual results if vocabulary tests were being used as part of a selection procedure for children too young to be tested with group scales. Norms must in fact be calculated separately for boys and girls.

Similar considerations apply to standardization for differences in age. Rough norms are not enough. Only if it can be shown that the adjustments do in fact yield scores which are independent, or almost independent, of differences in age and sex, can groups be compared which differ in age and sex composition. Matching becomes necessary, the very complication that norms are designed to avoid. We have tried to adopt a procedure which uses as much as possible of the information provided by the sample; and we feel that, in view of the purpose of the work, this needs no apology. In any case, more accurate norms, though much more troublesome to compute, are easy enough to use.

The first step in the procedure which we have adopted is to measure by how much the number of words correctly defined differs from the number expected for that sex and that age, the expected number being calculated from the fitted growth curves. A second step is then required. Variability also increases with advancing age; hence it is necessary to adjust the deviations from the expected scores so that these are made equivalent for children of different ages and incidentally for children of different sex. This can be done by multiplying the deviation at any observed age by the ratio of the standard deviation at that age to some fixed standard deviation. The method is perhaps a little cumbersome; but it has worked well with previous material, and, as will be shown, it also works well in the present instance. Nor does it seem likely that any rational transformation of the unit of measurement could make possible a simpler and more direct approach.

A sex difference in the variability of scores for intelligence tests has often been observed, and raises questions which have frequently been discussed, though certainly never settled. Thanks to its wide range of age, the present sample provides useful material for a further examination of this problem. The results will be presented in Section VII.

II. Increment of Mean Score with Age

In our previous paper [2] linear equations for increment of score with advancing C.A. were given by sexes for each vocabulary separately and for the four combined. The fit was good, without significant remainder, except for

W.I.S.C. (which in both sexes showed a decreasing rate of increment) and for Mill Hill A (which showed the same phenomenon, though, curiously enough, in girls only). Quadratics, however, gave an excellent fit in all three instances. Although the other remainders were not significant, nevertheless with M.H.A. for boys, M.H.B. for girls and with all four vocabularies combined for girls, quadratic equations improved the fit, the variance ratio for the addition of the quadratic term being significant in all three instances. Where a quadratic is used for one sex, it is desirable to use a second-order curve for the other sex as well, even though the improvement may be trivial; otherwise the sex difference at the extremes of the range of age may be exaggerated. For this reason, and for the sake of uniformity, quadratics have been used throughout. The importance of securing a good fit for the growth curves of the means, and the substantial improvement effected in some instances by the use of quadratics, is well brought out in the final testing of the norms by means of contingency tables, as described in Section V.

TABLE I. W.I.S.C. VOCABULARY
Quadratic Curves for Increment of Score. Analysis of Variance

Variation of Vocabulary Score	D. of F.	Sum of Squares (words)^2	Mean Square (words)^2	Variance Ratio	Signi-ficance
		BOYS			
Linear regression	1	15,858·3	15,858·3		
Addition of quadratic term	1	90·6	90·6	7·95	++
Remainder between arrays	17	288·0	16·94	1·50	—
Within arrays	960	10,839·3	11·29		
Total	979	27,076·2			
		GIRLS			
Linear regression	1	14,086·3	14,086·3		
Addition of quadratic term	1	206·6	206·6	23·34	++
Remainder between arrays	17	101·5	5·97	0·67	—
Within arrays	947	8,427·0	8·90		
Total	966	22,821·4			

The most important discrepancy to be corrected occurred with W.I.S.C.; the result of fitting quadratics is shown in Table I. Although M.H.A. is not used alone, mention may be made of the disappearance of the previously highly significant remainder shown by the girls. The variance ratio for the difference between the linear and curvilinear regressions, with degrees of freedom 1 and 964, is 18·63, and the variance ratio for the remainder, with degrees of freedom 17 and 947, is 1·13, as compared with the previous high figure of 2·10. The fitting of the growth curves for M.H.A. and M.H.B. combined is shown in Table II.

TABLE II. M.H.A.+M.H.B. VOCABULARIES
Quadratic Curves for Increment of Score. Analysis of Variance

Variation of Vocabulary Score	D. of F.	Sum of Squares (words)^2	Mean Square (words)^2	Variance Ratio	Significance
		BOYS			
Linear regression	1	77,810·3			
Addition of quadratic term	1	135·2	135·2	3·23	—
Remainder between arrays	17	906·0	53·29	1·28	—
Within arrays	960	39,996·3	41·66		
Total	979	118,847·8			
		GIRLS			
Linear regression	1	76,542·3			
Addition of quadratic term	1	394·2	394·2	12·10	++
Remainder between arrays	17	531·5	31·27	0·96	—
Within arrays	947	30,875·8	32·60		
Total	966	108,343·8			

The improvement with the four vocabularies combined is rather small. But the quadratic term is significant with the girls; and the variance ratio for the remainder is reduced from 1·20 to 0·86.

The regression equations used for calculating expected scores are as follows:

T.M.L.	Boys:	$y = -\ 0{\cdot}95 + 0{\cdot}0855267x + 0{\cdot}000216918x^2$,
	Girls:	$y = -\ 1{\cdot}82 + 0{\cdot}0976822x + 0{\cdot}000124580x^2$,
W.I.S.C.	Boys:	$y = -\ 1{\cdot}62 + 0{\cdot}189335x - 0{\cdot}000291713x^2$,
	Girls:	$y = -\ 3{\cdot}69 + 0{\cdot}215646x - 0{\cdot}000439474x^2$,
M.H.A.+B	Boys:	$y = -14{\cdot}18 + 0{\cdot}350433x - 0{\cdot}000356116x^2$,
	Girls:	$y = -18{\cdot}01 + 0{\cdot}404122x - 0{\cdot}000607071x^2$,
4 Combined	Boys:	$y = -16{\cdot}76 + 0{\cdot}625292x - 0{\cdot}000430884x^2$,
	Girls:	$y = -23{\cdot}52 + 0{\cdot}717450x - 0{\cdot}000921966x^2$,

where x = C.A. in completed months and y = score in words.

III. INCREMENT OF VARIANCE OF SCORE WITH AGE

The expected scores by age and sex having been obtained, the next step is to calculate multipliers for making deviations from expectation comparable over the whole range of age, five to fifteen years, and for children of opposite sex. Each multiplier is the ratio of the standard deviation at observed C.A. to some fixed standard deviation. In order to estimate the standard deviations at observed C.A. with as much accuracy as possible curves are fitted for the regression on C.A. of the variance of scores. This procedure was originally used by one of us in connection with norms for the Advanced Otis Scale (Roberts, Norman and Griffiths [5]), and has since been explained more fully (Roberts and Mellone [6].

Roberts [7]). It is a device that can be applied to any measurements when the range of age is great enough to yield a considerable increase in variance, and when no simple transformation of the units of measurement seems likely to provide a better alternative. For example, it has worked well for measurements of arterial blood pressure (Hamilton, Pickering, Roberts and Sowry [3]).

It was found that cubics are required. In four of the eight comparisons the addition of the cubic term effected significant improvement. What is more important is that straight lines or quadratics give expected variances for the youngest children which are far too low or even negative; this is shown clearly by Fig. 1.

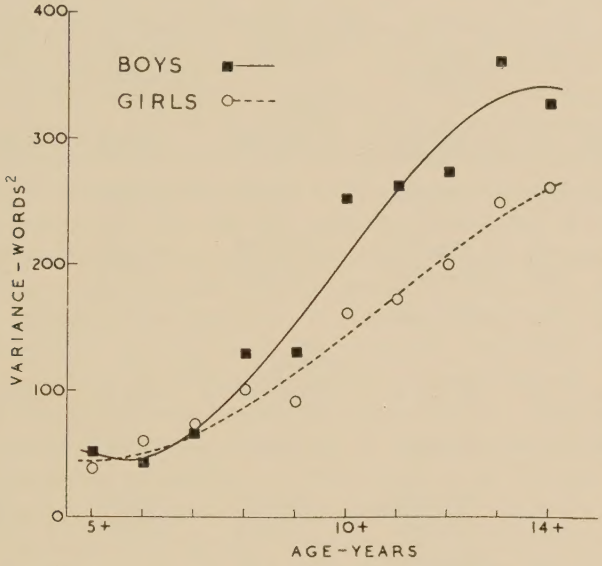

Fig. 1. Four vocabularies combined.
Increment of variance of score with age. Means of year age groups and fitted curves.

Table III. Regression of Variance of Score on Age
Significance of Remainders after Fitting Cubics

	Boys		Girls	
Vocabulary	Variance Ratio	Significance	Variance Ratio	Significance
T.M.L.	1·72	+	2·44	++
W.I.S.C.	2·02	++	1·36	–
M.H.A.+M.H.B.	1·94	+	0·61	–
4 combined	1·84	+	0·50	–

Degrees of freedom : 16 and ∞ throughout.

A summary of the results is given in Table III. With the girls the fit is excellent, except for T.M.L. With the boys the fit is not so good, all the remainders being significant, though not very large. It should be mentioned, however, that the final adjusted scores are a good deal less sensitive to changes in expected variances than they are to changes in expected means. Fig. 1, which shows the observed means of arrays and the fitted curves for all four vocabularies combined, brings out the closeness of the fit with the girls and some irregularity with the older boys. These features are also true of the individual vocabularies taken separately.

TABLE IV. FOUR VOCABULARIES COMBINED
Increment of Variance of Score with C.A. Analysis of Variance

Variation of Variance due to	D. of F.	Sum of Squares (words)^4	Mean Square (words)^4	Theoretical Variance (D. of F. = ∞)	Variance Ratio	Significance
			BOYS			
Linear Regression	1	11,508,784	11,508,784	58,342	197·26	++
Addition of quadratic term	1	40,055	40,055	93,885	0·43	—
Addition of cubic term	1	388,507	388,507	94,348	4·12	+
Remainder	16	2,832,823	177,051	96,023	1·84	+
Total	19	14,770,169				
			GIRLS			
Linear Regression	1	5,272,883	5,272,883	33,194	158·85	++
Addition of quadratic term	1	71,523	71,523	50,037	1·43	—
Addition of cubic term	1	103,369	103,369	50,537	2·05	—
Remainder	16	408,560	25,535	50,695	0·50	—
Total	19	5,856,335				

Table IV gives, as an example, the analysis of variance for the 4 vocabularies combined. The unit of grouping for C.A. is 6 months. The other analyses were fairly comparable.

The regression equations are as follows:

T.M.L. Boys: $y = 100{\cdot}88 - 3{\cdot}02508x + 0{\cdot}0286268x^2 - 0{\cdot}0000778981x^3$,
Girls: $y = 95{\cdot}31 - 2{\cdot}68802x + 0{\cdot}0239035x^2 - 0{\cdot}0000612672x^3$,

W.I.S.C. Boys: $y = 60{\cdot}88 - 1{\cdot}73925x + 0{\cdot}0166865x^2 - 0{\cdot}0000464019x^3$,
Girls: $y = 15{\cdot}23 - 0{\cdot}358567x + 0{\cdot}00380813x^2 - 0{\cdot}0000102841x^3$,

M.H.A. + B Boys: $y = 108{\cdot}20 - 3{\cdot}28393x + 0{\cdot}0335929x^2 - 0{\cdot}0000890677x^3$,
Girls: $y = 2{\cdot}85 - 0{\cdot}126640x + 0{\cdot}00404109x^2 - 0{\cdot}00000704577x^3$,

4 Combined Boys: $y = 751{\cdot}17 - 22{\cdot}7530x + 0{\cdot}223352x^2 - 0{\cdot}000609357x^3$,
Girls: $y = 252{\cdot}34 - 7{\cdot}28021x + 0{\cdot}0745683x^2 - 0{\cdot}000185767x^3$,

where x = age in completed months and y is variance of score in (words)^2.

IV. Norms

Each of the tables printed in the Appendix is divided into two parts. The first gives for each vocabulary, or group of vocabularies, C.A. in years and months, with the corresponding number of words; from these are read off the number of words by which the observed score exceeds or falls short of expectation. The second gives the multipliers which apply to successive 6-month age-groups. The observed deviation, plus or minus, is multiplied by the appropriate multiplier; and the final adjusted scores are thus made comparable for children of different ages and opposite sex.

The choice of the fixed standard deviations to which adjustment is made is explained in the notes appended to the tables of norms. All have a direct relation to the deviation from expectation in number of words as at some particular age within the range of the sample. We ourselves are using, with one exception, the plus or minus scores just as they stand. They have, of course, a mean of 0 for our sample. With the 4 vocabularies combined, however, a standard deviation of 15 is within the age range; so in this instance we add 100, thereby securing a frequency distribution with the conventional mean of 100 and standard deviation of 15.

One point in the calculation of the multipliers needs explanation. Some of the cubics are rather sharply inflected near the lower end of the range, which means that, if multipliers were given for each 6-month age-group, they would first of all increase rather rapidly from the group 5·0–5·5 before decreasing again. These tails are clearly rather unstable; so it seemed best to average the variances for successive groups, starting with the youngest, until the next older group showed a decrease from those younger groups so far combined. The six-month age-groups for multipliers may be thought rather coarse. With younger children the values are changing rapidly; with the 4 vocabularies combined the largest difference is 0·21, as the multiplier is 1·94 for boys of 7·0–7·5, and 1·73 for boys of 7·6–7·11. When the observed C.A. lies towards the end of a 6-month group, interpolation can be used for greater accuracy, though it seems hardly necessary. In the whole of the sample of 1,947 children, there are only 4 instances where interpolation gives a difference of as much as 2 points of score. To give an example, a girl aged 7·11 defined 66 words, 30 words above expectation; the application of the multiplier gives her an adjusted score of +54; and this is merely reduced to +52 on interpolation.

V. Efficiency of Adjusted Scores

It is desirable to see whether the adjusted scores are indeed achieving their purpose, that is, whether in fact they are independent of age and sex. Average allowance has been made for mean and variance; but there may be non-normal variations, for example skewness, which are different at different ages or in the two sexes. Accordingly, adjusted scores have been worked out for the 1,947 children making up the sample. Taking the sex comparisons first, the results are shown in Table V; no difference is significant; so that the adjusted scores are comparable for boys and girls.

TABLE V. ADJUSTED SCORES. COMPARISON OF SEXES

Vocabulary	Mean		S.D.	
	Boys	Girls	Boys	Girls
T.M.L.	+0·030	0·000	3·533	3·449
W.I.S.C.	−0·011	+0·021	3·002	3·007
M.H.A.+M.H.B.	−0·045	−0·051	5·803	5·716
4 Vocabularies	+0·006	−0·148	15·034	14·920

All differences non-significant

Independence of C.A. can be tested by forming contingency tables. In view of the results of Table V the scores of boys and girls can be added. The units of grouping are two years of age and 2 words, except for the 4 vocabularies combined, in which 5-word grouping is used. The results are shown in Table VI.

TABLE VI. TESTS OF INDEPENDENCE OF ADJUSTED SCORE AND C.A. : SUMMARY

Vocabulary	D. of F.	Chi Squared	*P*
T.M.L.	28	78·77	<0·00001
W.I.S.C.	28	44·23	0·03
M.H.A.+M.H.B.	52	54·23	0·39
4 Vocabularies Combined	52	52·13	0·47

It will be seen that M.H.A. + M.H.B. and the 4 vocabularies combined show almost perfect independence of C.A. and score. With W.I.S.C. chi-squared is significant, but not highly so. With T.M.L., on the other hand, there is a large and highly significant departure from independence. If the calculation is repeated, however, leaving out the youngest 2-year group, chi-squared falls to a value below the 5 per cent level of significance. The departure from independence, therefore, does not seem serious, as no one would want to use in isolation a single vocabulary of this kind for very young children.

As might be expected, the substitution of quadratic for linear growth curves for the means makes practically no difference when the improvement is small. Contingency tables based on the linear equations for T.M.L. and for the 4 vocabularies combined give values for chi-squared which are closely similar to those of Table VI. On the other hand, with M.H.A. and M.H.B. there was an appreciable falling off in rate of increment, which leads to a sharp increase in chi-squared from the 54·23 shown in Table VI to 82·24 if linear curves are used. No doubt W.I.S.C. would show the same effect; but we have not made this calculation. This illustrates the importance of securing a good fit and the inefficiency that would result from using approximate straight lines when a growth curve is appreciably curvilinear.

VI. Frequency Distribution of Vocabulary Score

With any intelligence scale designed to grade the whole population it is desirable that the scores should have an approximately normal frequency distribution. The adjusted scores for the 4 vocabularies combined give a single frequency distribution for the whole sample; they are no more than deviations from the regression line, further adjusted to equalize for differing variances at different ages and for different sexes. Departures from normality, such as skewness, will not be affected. Smoothed frequencies, with single words of score, as described in the preceding section must be used; and, although this may somewhat affect the degrees of freedom, the comparison should be broadly admissible.

Fig. 2. Four vocabularies combined.
Frequency distribution by 5 point groups of age, and sex adjusted scores, with fitted normal curve.

Fig. 2 shows the observed frequencies in 5-word groups, together with the fitted normal curve. Testing for normality by means of chi-squared, the central 16 5-word groups can be treated separately; the tails below -40 words and above $+39$ words, very nearly $\pm 2\frac{2}{3}\sigma$, must be lumped together to secure sufficient numbers. There are thus 18 arrays and, subject to the reservation made above about smoothing, 15 degrees of freedom. Chi-squared is 28·53, corresponding to a p of very nearly 0·02. The frequency distribution is thus approximately normal, and it therefore appears that scores on well constructed oral definition tests are as normally distributed as can reasonably be expected. The biggest contribution to chi-squared, 8·19, is due to an excess of scores above $+39$. The observed number is 16·75 against an expectation of 8·438. This is a feature of many intelligence tests; it is difficult to design a scale which does not allow a few very bright children to secure unduly high scores. The group $+15$ to $+19$

shows a deficiency; but the general feature that stands out in Fig. 2 is a slight tendency for the frequency curve to be flat-topped at about and rather below the mean. This phenomenon may be real, as it not infrequently appears with scales of different kinds.

VII. Further Examination of Variability of Score in Boys and Girls

The question of the relative variabilities of boys and girls in intelligence test scores has attracted interest over many years. The present sample has a wide range of age; and so is useful for examining the problem. The changing pattern as age increases is shown in Table VII. The means and standard

Table VII. Variability of Score

Age (Years)	Mean		S.D.		C.O.V. %		Ratio. Boys : Girls		
	Boys	Girls	Boys	Girls	Boys	Girls	Mean	S.D.	C.O.V.
5+	22·7	20·5	7·3	6·1	32·0	29·7	1·10	1·19	1·08
6+	29·5	27·0	6·5	7·7	21·9	28·6	1·09	0·84	0·76
7+	34·9	31·6	8·2	8·6	23·5	27·2	1·10	0·95	0·86
8+	42·8	39·7	11·4	10·1	26·6	25·4	1·08	1·13	1·05
9+	48·0	45·5	11·4	9·6	23·8	21·0	1·05	1·20	1·13
10+	57·0	53·5	15·9	12·7	27·9	23·8	1·07	1·25	1·17
11+	58·6	57·1	16·2	13·2	27·7	23·1	1·03	1·23	1·20
12+	67·7	64·2	16·6	14·2	24·5	22·1	1·06	1·17	1·11
13+	73·8	66·5	19·1	15·8	25·8	23·8	1·11	1·20	1·08
14+	77·7	73·4	18·2	16·2	23·4	22·1	1·06	1·12	1·06

All figures based on raw scores.

deviations are based on the simple raw scores by year age-groups for all four vocabularies combined. As was shown in the previous paper, boys have a higher mean score at all ages, though, as the eighth column shows, the relative excess tends to decline. With 6 and 7-year-old children the standard deviation is higher in girls, but thereafter is considerably higher in boys. The coefficient of variation gives the clearest picture. The ratios shown in the last column reveal a clear trend. In the youngest age-group, 5·0–5·11, boys have the higher C.O.V.; but it would be unsafe to deduce too much from tests given to such young children, In the seventh year girls are much more variable. At about 8 or 9 years equality is reached; and thereafter the variability of boys relative to that of girls increases steadily to about 11 or 12 years; it then declines just as steadily to age 15.

The ratios of the C.O.V.'s are shown in Fig. 3, where the figures of the last column of Table VII have been plotted. A continuous line is shown which represents the ratio of the C.O.V.'s when calculated from the curves fitted for increment with age of mean and variance. The trends are smoothed; but in fact little is added to the clear evidence of the figures based on the simple raw scores.

It is tempting to conclude that, in relative variability, vocabulary tests may be typical of intelligence scales in general. If so, the greater variability of boys may be a passing phase in development, absent in the young child, reaching a maximum at 11 or 12 years, and perhaps (though this is an extrapolation) disappearing by the time adult life is reached. The phenomenon of greater male variability is particularly striking, because most of the evidence comes from group tests administered at about the age when the difference is maximal. All this is somewhat speculative; but we hope that our data make some contribution, and that others will provide further evidence based on other intelligence scales. The problem is important. If indeed the very real greater male variability is a passing phase, then allowance is rightly made by scoring boys and girls separately in scholarship examinations, and basing grammar school places, for example, on an equal proportion of each sex, and not on a dividing line of score applied to both boys and girls.

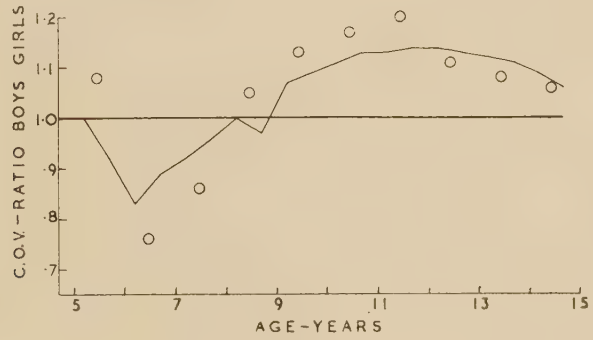

Fig. 3. Four vocabularies combined.
Ratio of coefficients of variation (boys/girls). Circles based on raw scores by year age groups ; continuous line based on fitted curves for increment of mean and variance with age.

VIII. Summary

1. A 3 per cent sample of the school children of the City and County of Bristol was selected by visiting all the schools, as well as a number outside, and testing children between the ages of 5·0 to 14·11 years whose homes were within the city and whose birthdays fell on the first day of any calendar month. Four vocabulary scales were used: that from the Terman–Merrill Scale (Form L), Mill Hill Vocabularies A and B, and that from the Wechsler Intelligence Scale for Children.

2. The present paper gives norms for the vocabularies from the Terman–Merrill and Wechsler Intelligence Scales, the oral definitions form of Mill Hill Vocabulary Scales A and B, and for all four vocabularies combined. In view of the sex difference, norms are given separately for boys and girls.

3. Curves have been fitted for the increment of score with advancing age; and tables are given from which deviations from expectation can be read off.

4. Further curves have been fitted for the increment of variance with advancing age. Deviations from expectation can be made comparable for children of different ages by multiplying by the ratio of the standard deviation at observed age to a fixed standard deviation. Tables of multipliers are given for making this further adjustment, which also makes scores comparable for children of opposite sex.

5. It is shown that for this sample the adjusted scores are, with small exceptions, independent of age at test, and that they equalize for children of opposite sex.

6. The frequency distribution is substantially normal.

7. In terms of the coefficient of variation girls were more variable than boys in the younger age-groups. At about 8 or 9 years equality is reached; and thereafter boys become progressively more variable than girls up to the age of 11 or 12. Beyond this age the difference between the sexes is progressively reduced.

Acknowledgments

We are most grateful to Professor Lewis M. Terman and to the Houghton Mifflin Company for permission to use the Vocabulary test from the Terman–Merrill Intelligence Scale [8]; to Mr. J. C. Raven [4] for permission to use the Oral Definitions Form of his Mill Hill Vocabulary Scale; and to Dr. David Wechsler [9] for permission to use the Vocabulary from his Intelligence Scale for Children. Special thanks are due to the Bristol Education Committee and to the Chief Education Officer, Mr. G. H. Sylvester, who gave permission for work to be carried out in their schools, and in particular to the Headmasters and Headmistresses both of the Local Education Authority's schools and of independent schools, for their willing and active co-operation.

APPENDIX

Notes to Table VIII

1. To obtain scores adjusted for differences in C.A. and sex, read off expected number of words for observed C.A. Multiply the difference (plus or minus) between expected and obtained number of words by multiplier corresponding to observed C.A.

2. For 4 vocabularies, combined scores are adjusted to a standard deviation of 15, which is that shown by boys of about 10 years 8 months and by girls of about 13 years 0 months. Mean of zero can be changed to 100 by adding 100 algebraically to the adjusted score.

3. For M.H.A. + M.H.B. scores are adjusted to a standard deviation of 5·7151, which is the root mean variance of girls in the experimental sample, and corresponds to a C.A. of about 9 years 0 months for boys and about 9 years 9 months for girls.

4. For T.M.L. the scores are adjusted to a standard deviation of 3·4357, which is the root mean variance of girls as before.

5. For W.I.S.C. the scores are adjusted to a standard deviation of 2·9851, which is the root mean variance of girls as before.

TABLE VIII. NORMS ; EXPECTED SCORES AND MULTIPLIERS

T.M.L. C.A. Boys	T.M.L. C.A. Girls	T.M.L. Exp'd No. of Words	W.I.S.C. C.A. Boys	W.I.S.C. C.A. Girls	W.I.S.C. Exp'd No. of Words
				5·0– 5·5	8
	5·0	4	5·0– 5·5	5·6– 5·11	9
5·0 – 5·4	5·1 – 5·8	5	5·6 – 6·0	6·0– 6·6	10
5·5 – 6·1	5·9 – 6·5	6	6·1 – 6·6	6·7– 7·1	11
6·2 – 6·9	6·6 – 7·2	7	6·7 – 7·1	7·2– 7·8	12
6·10– 7·5	7·3 – 7·10	8	7·2 – 7·9	7·9– 8·4	13
7·6 – 8·1	7·11– 8·6	9	7·10– 8·4	8·5– 9·0	14
8·2 – 8·9	8·7 – 9·2	10	8·5 – 9·0	9·1– 9·8	15
8·10– 9·5	9·3 – 9·10	11	9·1 – 9·8	9·9– 10·5	16
9·6 –10·0	9·11–10·6	12	9·9 –10·5	10·6–11·3	17
10·1 –10·7	10·7 –11·1	13	10·6 –11·1	11·4–12·2	18
10·8 –11·2	11·2 –11·9	14	11·2 –11·11	12·3–13·3	19
11·3 –11·9	11·10–12·5	15	12·0 –12·8	13·4–14·5	20
11·10–12·4	12·6 –13·0	16	12·9 –13·7	14·6–14·11	21
12·5 –12·10	13·1 –13·7	17	13·8 –14·6		22
12·11–13·5	13·8 –14·2	18	14·7 –14·11		23
13·6 –13·11	14·3 –14·9	19			
14·0 –14·6	14·10–14·11	20			
14·7 –14·11		21			

MULTIPLIERS

C.A.	T.M.L. Boys	T.M.L. Girls	W.I.S.C. Boys	W.I.S.C. Girls
5·0– 5·5	2·03	2·00	1·31	1·31
5·6– 5·11				1·31
6·0– 6·5				1·28
6·6– 6·11				1·25
7·0– 7·5	1·99		1·30	1·21
7·6– 7·11	1·65	2·00	1·21	1·17
8·0– 8·5	1·38	1·74	1·12	1·13
8·6– 8·11	1·18	1·46	1·03	1·08
9·0– 9·5	1·03	1·26	0·95	1·04
9·6– 9·11	0·92	1·10	0·89	1·00
10·0–10·5	0·84	0·99	0·83	0·97
10·6–10·11	0·78	0·90	0·79	0·93
11·0–11·5	0·73	0·83	0·75	0·91
11·6–11·11	0·69	0·78	0·72	0·88
12·0–12·5	0·66	0·73	0·70	0·86
12·6–12·11	0·64	0·70	0·68	0·84
13·0–13·5	0·62	0·68	0·67	0·82
13·6–13·11	0·62	0·66	0·67	0·81
14·0–14·5	0·62	0·65	0·68	0·80
14·6–14·11	0·62	0·64	0·69	0·80

M.H.A.+M.H.B.

C.A. Boys	C.A. Girls	Exp'd No. of Words	C.A. Boys	C.A. Girls	Exp'd No. of Words
	5·0 – 5·1	4			
	5·2 – 5·4	5	10·7 –10·10	10·11–11·3	25
5·0 – 5·3	5·5 – 5·7	6	10·11–11·2	11·4 –11·7	26
5·4 – 5·6	5·8 – 5·10	7	11·3 –11·6	11·8 –11·11	27
5·7 – 5·9	5·11– 6·1	8	11·7 –11·10	12·0 –12·3	28
5·10– 6·1	6·2 – 6·4	9	11·11–12·2	12·4 –12·8	29
6·2 – 6·4	6·5 – 6·8	10	12·3 –12·6	12·9 –13·1	30
6·5 – 6·7	6·9 – 6·11	11	12·7 –12·10	13·2 –13·5	31
6·8 – 6·11	7·0 – 7·2	12	12·11–13·2	13·6 –13·10	32
7·0 – 7·2	7·3 – 7·6	13	13·3 – 13·7	13·11–14·3	33
7·3 – 7·6	7·7 – 7·9	14	13·8 –13·11	14·4 –14·8	34
7·7 – 7·9	7·10– 8·1	15	14·0 –14·3	14·9 –14·11	35
7·10– 8·1	8·2 – 8·4	16	14·4 –14·8		36
8·2 – 8·4	8·5 – 8·8	17	14·9 –14·11		37
8·5 – 8·8	8·9 – 8·11	18			
8·9 – 8·11	9·0 – 9·3	19			
9·0 – 9·3	9·4 – 9·7	20			
9·4 – 9·7	9·8 – 9·11	21			
9·8 – 9·11	10·0 –10·2	22			
10·0 –10·2	10·3 –10·6	23			
10·3 –10·6	10·7 –10·10	24			

MULTIPLIERS

M.H.A.+M.H.B.

C.A.	Boys	Girls	C.A.	Boys	Girls
5·0–5·5	1·63	1·90	10·0–10·5	0·84	0·97
5·6–5·11		1·73	10·6–10·11	0·79	0·92
6·0–6·5	1·57	1·59	11·0–11·5	0·75	0·88
6·6–6·11	1·47	1·47	11·6–11·11	0·72	0·85
7·0–7·5	1·35	1·36	12·0–12·5	0·69	0·82
7·6–7·11	1·24	1·27	12·6–12·11	0·67	0·79
8·0–8·5	1·13	1·20	13·0–13·5	0·65	0·76
8·6–8·11	1·04	1·13	13·6–13·11	0·64	0·74
9·0–9·5	0·96	1·07	14·0–14·5	0·63	0·72
9·6–9·11	0·90	1·01	14·6–14·11	0·62	0·69

FOUR VOCABULARIES COMBINED

C.A. Boys	C.A. Girls	Exp'd No. of Words	C.A. Boys	C.A. Girls	Exp'd No. of Words
			9·8	10·1 –10·2	50
	5·0	16	9·9 – 9·10	10·3 –10·4	51
	5·1 – 5·2	17	9·11–10·0	10·5 –10·6	52
	5·3	18	10·1 –10·2	10·7 –10·8	53
5·0	5·4 – 5·5	19	10·3 –10·4	10·9 –10·10	54
5·1 –5·2	5·6 – 5·7	20	10·5 –10·6	10·11–11·0	55
5·3 –5·4	5·8 –	21	10·7 –10·8	11·1 –11·2	56
5·5	5·9 – 5·10	22	10·9 –10·10	11·3 –11·5	57
5·6 –5·7	5·11– 6·0	23	10·11–11·0	11·6 –11·7	58
5·8 –5·9	6·1	24	11·1 –11·2	11·8 –11·9	59
5·10–5·11	6·2 – 6·3	25	11·3 –11·4	11·10–11·11	60
6·0	6·4 – 6·5	26	11·5 –11·6	12·0 –12·1	61
6·1 –6·2	6·6 – 6·7	27	11·7 –11·8	12·2 –12·4	62
6·3 –6·4	6·8	28	11·9 –11·10	12·5 –12·6	63
6·5 –6·6	6·9 – 6·10	29	11·11–12·0	12·7 –12·8	64
6·7	6·11– 7·0	30	12·1 –12·2	12·9 –12·10	65
6·8 –6·9	7·1 – 7·2	31	12·3 –12·4	12·11–13·1	66
6·10–6·11	7·3 – 7·4	32	12·5 –12·6	13·2 –13·3	67
7·0 –7·1	7·5	33	12·7 –12·8	13·4 –13·5	68
7·2 –7·3	7·6 – 7·7	34	12·9 –12·10	13·6 –13·8	69
7·4 –7·5	7·8 – 7·9	35	12·11–13·0	13·9 –13·10	70
7·6	7·10– 7·11	36	13·1 –13·2	13·11–14·1	71
7·7 –7·8	8·0 – 8·1	37	13·3 –13·4	14·2 –14·3	72
7·9 –7·10	8·2 – 8·3	38	13·5 –13·6	14·4 –14·6	73
7·11–8·0	8·4	39	13·7 –13·8	14·7 –14·8	74
8·1 –8·2	8·5 – 8·6	40	13·9 –13·10	14·9 –14·11	75
8·3 –8·4	8·7 – 8·8	41	13·11–14·0		76
8·5	8·9 – 8·10	42	14·1 –14·2		77
8·6 –8·7	8·11– 9·0	43	14·3 –14·4		78
8·8 –8·9	9·1 – 9·2	44	14·5 –14·7		79
8·10–8·11	9·3 – 9·4	45	14·8 –14·9		80
9·0 –9·1	9·5 – 9·6	46	14·10–14·11		81
9·2 –9·3	9·7 – 9·8	47			
9·4 –9·5	9·9 – 9·10	48			
9·6 –9·7	9·11–10·0	49			

MULTIPLIERS : FOUR VOCABULARIES COMBINED

C.A.	Boys	Girls
5·0– 5·5	2·19	2·28
5·6– 5·11		2·27
6·0– 6·5		2·19
6·6– 6·11	2·14	2·07
7·0– 7·5	1·94	1·93
7·6– 7·11	1·73	1·79
8·0– 8·5	1·54	1·66
8·6– 8·11	1·39	1·55
9·0– 9·5	1·26	1·45
9·6– 9·11	1·16	1·36
10·0–10·5	1·07	1·28
10·6–10·11	1·01	1·21
11·0–11·5	0·95	1·15
11·6–11·11	0·91	1·10
12·0–12·5	0·87	1·06
12·6–12·11	0·85	1·02
13·0–13·5	0·83	0·98
13·6–13·11	0·81	0·96
14·0–14·5	0·81	0·93
14·6–14·11	0·81	0·91

REFERENCES

[1] DUNSDON, M. I., and ROBERTS, J. A. F. (1953). ' The relation of the Terman–Merrill vocabulary test to mental age in a sample of English children.' *Brit. J. Stat. Psychol.*, VI, 61–70.

[2] DUNSDON, M. I., and ROBERTS, J. A. F. (1955). ' A study of the performance of 2,000 children on four vocabulary tests.' *Brit. Stat. J. Psychol.*, VIII, 3–15.

[3] HAMILTON, M., PICKERING, G. W., ROBERTS, J. A. F., and SOWERBY, G. S. G. (1954). ' The aetiology of essential hypertension : (2) Scores for arterial blood pressures adjusted for differences in age and sex.' *Clin. Sci.*, XIII, 37–49.

[4] RAVEN, J. C. (1948). *The Mill Hill Vocabulary Scale.* London : H. K. Lewis.

[5] ROBERTS, J. A. F., NORMAN, R. M., and GRIFFITHS, R. (1935). ' Studies on a child Population : (i) definition of the sample method of ascertainment, and analysis of the results of a group intelligence test.' *Annals of Eugenics*, VI, 319–338.

[6] ROBERTS, J. A. F., and MELLONE, M. A. (1952). ' On the adjustment of Terman–Merrill I.Q.s to secure comparability at different ages.' *Brit. Stat. J. Psychol.*, V, 65–79.

[7] ROBERTS, J. A. F. (1953). ' The use of regressions involving variances of dependent variates for calculating age corrected scores.' *Biometrics*, IX, 267.

[8] TERMAN, L. M., and MERRILL, M. A. (1937). *Measuring Intelligence.* London : Harrap.

[9] WECHSLER, D. (1949). *Intelligence Scale for Children.* New York : The Psychological Corporation.

[10] YATES, A. J. (1956). ' The use of vocabulary in the measurement of intellectual deterioration : A review.' *J. Ment. Sci.*, CII, 409–440.

Vol. X
Part I

The British Journal of Statistical Psychology

May
1957

PSYCHOLOGICAL FACTORS: THEIR NUMBER, NATURE, AND IDENTIFICATION

By G. Bernyer
Centre National de la Recherche Scientifique, Paris

Summary

The following paper presents a brief survey of current factorial procedures, with a discussion of their chief advantages and limitations. It is intended primarily for research workers in general psychology who desire to analyse the factors involved in their data without making a detailed study of all available procedures. An attempt is made to determine the commoner faults, either of theory or of practice, which have rendered so many recent factorial investigations contradictory or inconclusive; and a number of defects are noted in existing techniques which appear to call for further investigation by the mathematical theorist. The main conclusion is that, like other statistical procedures, factor analysis is merely an adjunct to psychological research, not a method sufficient in itself[1].

The Choice of Procedures

During the past fifty years, factor analysis has undergone a rapid development and provoked a series of perplexing controversies. The literature, even if we include only publications dealing with factorial procedures and disregard those summarizing actual results, is now so vast that the investigator finds himself confronted by a formidable mass of contributions, which recommend widely different techniques and often lead to conclusions that are entirely contradictory. The psychologist who does not wish to embark on a comprehensive study of the whole subject, and wants merely to select the method most appropriate to his own immediate problem, is faced with a bewildering task if he is to appreciate the technical difficulties, the psychological implications, and the special limitations of the particular procedure that he has decided to apply. It may therefore be helpful to attempt a brief review of the practical questions encountered by the ordinary worker who desires to choose and apply some suitable technique. I shall confine my remarks chiefly to investigations involving tests of ability or aptitude, and shall deal chiefly with the two methods that are most frequently used by the psychologist, namely, the method of general and group factor analysis on the one hand, and the method of multifactorial analysis (including rotation to simple structure) on the other.

[1] *Editorial Note*—Both the author and the editor are indebted to the authorities of the *Centre National de la Recherche Scientifique* for permission to publish this paper here. The French version was originally presented at the International Colloquium on Factor Analysis, arranged by the *Centre* and published, with a full account of the discussion that followed, in the Report of their proceedings (Paris, 1955). Many requests have already been received for a copy to be available in English. We are grateful to Mlle. G. Régnier for undertaking the translation, and to Dr. Charlotte Banks for revising the manuscript. Students and others can obtain reprints at cost price by writing to the Assistant Editor, Dr. Charlotte Banks, University College, London.

Fundamental Assumptions

Both procedures start with the broad hypothesis that the marks obtained by a particular examinee in a particular test may, to a first approximation, be treated as a linear function of certain hypothetical variables or 'factors'; and both recognize two types of factor, namely, (*a*) 'common factors', i.e. those that appear either in all the tests (the general factor) or in some of the tests (group factors), and (*b*) 'specific factors', i.e. those peculiar to some single test alone. All such factors, except perhaps those that are invoked in the later stages of multifactorial analysis, are assumed to be statistically independent.

Here it is too often forgotten that, even when we restrict ourselves to this apparently simple hypothesis, the problem of determining the factors still possesses an infinity of solutions. To meet this difficulty and limit the number of solutions available, most writers commonly invoke a 'principle of economy', very much the same in both procedures. This principle requires us to extract a minimum number of common factors. By thus minimizing the common factor variance, however, we maximize the specific factor variance. Moreover, if each of the two alternative procedures take as their point of departure one and the same bipolar analysis, the results turn out to be statistically equivalent; we can pass from one set to the other by an appropriate rotation. On the other hand, the factors actually furnished by the two methods of analysis differ widely from each other in consequence of the further postulates which each imposes in order to reach the final solution.

Analysis by General and Group Factors

The 'two-factor theory' of Spearman [1] sought to interpret the marks obtained in a test by each examinee by means of a single general factor only, together with a factor specific to the test. This seemed to provide a solution of the factorial problem which was attractive at once by its neatness and by its simplicity. But, as soon as it was fully realized that this simple assumption fails to fit the psychological data [2], and that a single general factor, g, is insufficient to explain all the correlations between the tests, it became necessary to reintroduce group factors as well [3]. Spearman himself for long refused to admit the need for such supplementary factors; and, when at length he agreed to adopt them, he again applied a strict principle of economy, and insisted that the group factors should be as few as possible and that their existence and statistical significance should first be fully established.

With a view to determining their number and extent, he himself relied on the examination of tetrad differences [4]. In its practical application, however, this procedure presents a number of grave difficulties. It is extremely laborious; it still does not always lead to a unique factorial solution; and the sampling errors for the tetrad differences can only be calculated by the aid of formulae which seem in point of fact to be somewhat crude approximations. In practice, therefore, the employment of this procedure has remained confined to small tables containing only a few tests and including only a single group factor—a limitation that has not always been noticed.

An alternative method, more recent, which has been largely overlooked by psychologists, could be substituted with great advantage for that employed by Spearman. This is the method proposed by P. Delaporte [5]. It permits much larger batteries to be analysed, and allows us to use tables giving the confidence intervals for the ratio between the correlation coefficients observed. It thus permits the investigator to establish the factorial pattern quite quickly and without detailed calculation. However, in its practical application a few minor questions still remain to be solved: in certain cases it is by no means easy to decide whether the principle of economy should be applied to the number or to the extent of the factors; and consequently a subjective choice, which takes into account the actual content of the test, will then be necessary.

Very few psychologists have in point of fact continued to follow the line proposed by Spearman. Burt has reviewed the reasons which led him to criticize this simplified method of analysis [6]. With the somewhat drastic principle of economy advocated by Spearman the chief difficulty is that the general factor which is actually extracted can vary widely as a result of the modifications that may be introduced in the composition of the test-battery. In the majority of cases its effect is to suppress the broadest group factor, which gets incorporated into the general factor, and as a result imparts to the modified general factor an additional content which should properly be assigned to the group factor, for example, a verbal content or a spatial content according to the nature of the tests employed. As a result the general factors so obtained cease to be comparable from one battery to another.

To overcome this difficulty, several investigators have adopted the bi-factorial method of Holzinger [7] which interprets the intercorrelations of the tests in terms of a general factor, *g*, and a number of group factors which are mutually exclusive. However, the procedures proposed for deciding how the tests are to be grouped, for calculating the factor saturations, and for testing the statistical significance of the residual correlations left after the common factors have been extracted, appear for the most part to be rather crude.

In a recent paper on group factor analysis [8] Burt has furnished a method of great value for those psychologists who wish to carry out a group factor analysis on a more precise basis. Readers will there find, clearly summarized, a critical study of the various procedures most commonly used for establishing the factor pattern (subjective classification, the B coefficient of Holzinger, Tryon's method of profiles, and the device of intercolumnar correlations), together with an account of what Burt calls the 'method of summation' for calculating the factor saturations illustrated by worked examples. Moreover, the procedure described indicates how a preliminary bipolar analysis may be used with advantage to determine the best grouping of the tests, and also allows the investigator to calculate factor saturations even when there is a partial overlap between certain of the group factors.

Subdivided Factors

It is likewise to Burt that we owe an important extension of the group factor procedure, namely, the method of 'subdivided factors' [9]. After describing

the hypothesis put forward by McDougall and other writers according to which the structure of the mind is believed to be essentially 'hierarchical', Burt shows that a divergent scheme of classification often proves to be applicable to psychological data, and that, when a table of correlations fulfils certain conditions (in particular when the residual cross-correlations remaining after a bipolar factor has been extracted are non-significant), it is possible to demonstrate over and above the general factor certain broad group factors (e.g. an intellectual factor and a practical factor) which in turn subdivide themselves into narrower group factors (e.g. a verbal factor, a numerical factor, and the like, subsumed under the heading of the intellectual group factor) and such that these in turn can be further subdivided into factors which are narrower still.

Many British authors (e.g. Vernon) have successfully used factor patterns of this type. The preliminary bipolar analysis serves as a point of departure for the calculation of the subdivided group factors. And with this mode of analysis not only the arrangement, but also the number of factors extracted (which is somewhat greater than that of the corresponding bipolar factors obtained with the classical procedures) is automatically determined by the number of factors extracted in the preliminary bipolar analysis. Thus, if we take a common instance, namely, a bipolar analysis involving four factors—one general and positive, one bipolar dividing this into two, and two more subdividing its positive and negative sections into two subsections each, with cross-correlations of approximately zero, we find that, when expressed in terms of a group-factor classification with positive saturations only, this yields as many as seven subdivided factors—one general and a pair of broad group factors, each subdividing into a different pair of narrower factors.

Although group factor analysis has been used fairly widely in psychology, the procedure has its limitations. In the first place, if the preliminary method of calculation is based on the method of simple summation, which is ordinarily substituted for the more exact technique of weighted summation, then the more rigorous tests for statistical significance, which are appropriate to weighted summation, are no longer strictly valid, and must be regarded as approximate only. In the second place, group factor analysis, like every other factorial procedure, can only be applied to batteries of tests which have been carefully planned in view of the problem to be investigated. "The common idea", says Burt, "that factor analysis can be mechanically applied to any set of data that happens to be handy, and that factors can thus be 'discovered', is entirely misleading."

Multifactorial Analysis

Thurstone, in opposition to the theory put forward by Spearman, starts from the hypothesis that the inter-correlations of a battery of tests are to be explained, not in terms of a single general factor, but by a small number of common factors [10, 11]. As is well known, the proof that all the tetrad differences (i.e. minor determinants of order two) are equal to zero is the mathematician's

regular way of demonstrating that the rank of the matrix from which they are derived is one. Thurstone invokes the general form of this theorem. "In the multifactorial problem", he says, "we inquire at the outset what is the rank of the given table of intercorrelations; and this rank indicates the number of factors which have to be postulated in order to interpret the correlations experimentally obtained." However, in practice "the rank of a correlation matrix is always equal to its order, because of the accidental errors of sampling entering into each of its elements. Consequently, we are compelled to make use of an approximation to the rank of the matrix. We ask, in point of fact, whether it is possible to construct a theoretical correlation matrix of lower rank, which will only differ to a slight extent from the correlations experimentally observed. When such a correlation matrix can be found, then its rank gives the number of factors which have to be postulated to interpret the observed correlations" [12].

The method thus proposed by Thurstone for solving this problem appears to involve several difficulties, which I shall now briefly attempt to indicate.

Communalities

The first question to be settled is the size of what are called the 'communalities', that is to say, the quantities which are to be entered in the principal diagonal of the correlation matrix in order to make its rank a minimum. Now, at the beginning of the analysis, the rank of a correlation matrix is unknown; and strictly speaking we can only secure estimates for the communalities by determining (as in the case of a correlation matrix of rank one) the lowest order of minor determinants which are equal to zero. Several working procedures, all of them more or less crude, have been suggested and discussed. But the majority of analysts appear content to adopt Thurstone's short cut, and, before extracting each factor, insert in each diagonal cell the largest correlation in the column. When the number of tests is large, the error thus introduced is comparatively unimportant. However, Burt and other English investigators prefer to carry out the estimation of the communalities or self-correlations by successive approximation. With smaller tables of correlations some investigators may be sufficiently interested to try the method of maximum likelihood described by Lawley, provided they are ready to face the rather laborious calculations involved. This yields efficient estimates, and incidentally provides a criterion for deciding the number of factors to be extracted [13 and refs.].

The next problem consists in computing from the table of correlations as thus completed the various common factors which will account for the observed coefficients within the margin allowed by the fluctuations of sampling. Thurstone seems now [14] to view with more favour than before [11] the method of 'principal components' as described by Hotelling: this is the procedure with which British writers are more familiar as the 'method of principal axes', originally described by Karl Pearson [15]. With this procedure we obtain a series of orthogonal factors by "extracting at each stage that particular factor which will account for a maximum amount of the common factor variance"—the so-called method of

'weighted summation' [16]. The method of 'simple summation' described by Burt and the equivalent procedure called the 'centroid method' by Thurstone are quicker to apply, and furnish a reasonably good approximation to the principal components; the centroid method is that most commonly employed by American writers. All these procedures can readily be applied to practically any table of correlations. The chief difficulty that remains is to determine at what point the analysis must be stopped.

Number of Factors

How many factors should be extracted has been a question on which numerous inquiries have been published, and is still far from being completely solved. Many criteria have been put forward for testing the statistical significance of the residual correlations which remain after a given number of common factors have been extracted [17]. Of these the more exact demand lengthy calculations and moreover are not valid when the analysis has been carried out by the centroid procedure. The majority are based on more or less empirical rules, which in point of fact scarcely yield any more information than can be obtained by simple inspection without any calculations derived from the residuals. Indeed, in practice, many factorists employ none at all. Thurstone and his pupils apparently prefer to extract more factors than is stictly necessary, and then at the end of the analysis to discard those which are not susceptible of psychological interpretation. This lack of rigour is perhaps partly justified, since, as Thomson has indicated, the sampling errors can themselves produce common factors which will not be concentrated in the last centroid factor, but will be mingled with all such factors; and Burt has further insisted that "the discovery that, with the sample used, a given factor turns out to be non-significant cannot be taken to prove that it is non-existent".

Actually the aim of the psychologist is not so much to extract from a correlation matrix the minimum number of factors as to verify the existence of the factors that he has postulated, and if necessary to discover whether there may not be further groupings of the tests employed which have hitherto been overlooked. In order to judge whether the number of factors extracted is sufficient, he will in the end nearly always need to examine the particular results obtained. For example, in an analysis where three factors furnished non-significant residuals as judged by McNemar's criterion, I myself have been able to show, simply by examining the diagram obtained on plotting the test-vectors, that four factors at least would be required to group the tests correctly and achieve a simple structure [18].

The greatest difficulty that besets the use of factor analysis springs from the fact that the 'components' or 'mental factors' that such investigations continue to reveal do not form a small or sharply limited group. On the contrary, as the facts seem clearly to prove, the progress of factorial work leads to an ever increasing number. In a recent monograph dealing with the description of aptitude and achievement tests in terms of rotated factors [19], J. W. French has enumerated

well over sixty. Now in any given factorial investigation it is possible to extract only a very few fully significant factors. The actual number is unavoidably limited by the number of the subjects who have been tested. After a study of a series of different researches Burt considers that, as a rule, with a battery of about ten to twenty tests, at least twenty subjects are needed to establish one factor, at least fifty subjects to establish two, over a hundred to establish three, and as many as two or three hundred to establish four. Thus a factorial analysis can achieve a result which is at once statistically accurate and capable of a clear interpretation only when the number of factors aimed at has been scrupulously limited from the very outset.

Simple Structure

The different methods of analysis which we have briefly enumerated—the method of simple summation or centroid analysis, the method of principal components or weighted summation, the method of maximum likelihood—yield factors which have negative loadings in certain tests and which often vary widely if the composition of the test-battery is changed. Such factors, so Thurstone contends, must be devoid of 'scientific meaning'. They do not permit us to "interpret the various tests as functions of the mental aptitudes which those tests elicit". In order to acquire a meaning, therefore, the factors should be subjected to rotations, which are continued until the factorial matrix has the largest possible number of values which are either zero or approximately zero. If the investigator succeeds by this device in obtaining at least one zero saturation in each row, and at least as many zeros in each column as there are factors, he is said to have achieved a 'simple structure'. When such a structure has been secured, however, the group factors that emerge are no longer of necessity orthogonal to each other. Thurstone regards them as 'primary factors', and believes them to be invariant and susceptible of psychological interpretation [14].

In practice, the search for such a solution is not without difficulties of its own. No method that is absolutely automatic and objective has been devised for securing it. And the successive rotations which are needed require long and laborious trials. In practice, since there are no formulae for determining the errors in the factor saturations thus obtained, many analysts have arbitrarily adopted their own limit of significance; and indeed often seem a little too prone to consider that they have satisfied the criteria for a simple structure. By such procedures they manage to start from almost any kind of correlation table and reach a set of 'primary factors' whose stability and whose psychological interest are in point of fact far from being adequately established. Here assuredly there lies a great danger—one in fact on which Thurstone himself has insisted. "Factor analysis", he says, "is a scientific method which has to be adapted to each problem: it is no mere statistical procedure; and it is not a routine device which one can fruitfully apply to any table whatsoever." It would seem, then, that every factorial investigation ought to start with a plan of inquiry, and that a 'simple structure' can have no definite meaning unless it is reached as a confirmation of factors postulated at the outset.

The Identification and Invariance of Factors

By way of defending the superiority of their own particular method, the champions of different factorial procedures commonly assert that theirs will yield factors which can be most readily identified and which will remain the most stable when one changes either the battery of tests or the population tested. And in order to secure an objective basis for determining the identity of the factors thus extracted, a number of widely different procedures have been devised. The following seem to be the most satisfactory suggestions [20]: (i) when several wholly independent analyses have been carried out, to compare the amount contributed to the variance by corresponding factors in different analyses; (ii) when the experiments have been carried out on the same group of subjects, to calculate the correlations between the estimates for corresponding factors; (iii) when the experiments have been carried out with the same battery of tests, to compare the factor saturations for the several tests by means of an appropriate coefficient: for this purpose Burt has devised a 'coefficient of similarity' between the two corresponding sets of saturations, while Cattell prefers to use a 'marker criterion' or a 'coefficient for similarity of patterns'.

Hitherto these procedures have been but rarely used, and that for several reasons. Until quite recently there have been comparatively few factorial investigations whose results would permit their application: it is in fact rather exceptional for the factor measurements to be estimated in detail for each of the persons tested; and, even so, a knowledge of the reliability of the measurements would seem to be necessary before any trustworthy information can be deduced from their correlations. Moreover, it is seldom possible to compute even the simpler type of coefficient based on a comparison of the factor saturations, since, in their desire to carry out original work, the majority of authors appear reluctant to adopt batteries that have already been used in previous researches, and nearly always prefer to substitute new tests of their own. French observes that, in the comparative study that he has made of some 69 analyses, he has been obliged to limit himself to a subjective comparison of the tests because there were only four instances in which precisely the same batteries had been employed [19]. Moreover, the practical value of the different procedures still needs to be more fully demonstrated. Leyden [21], using data from a battery of tests administered to the same group of pupils at intervals of 1 to 3 years, has applied most of the foregoing criteria to test the stability of the general factor, and finds that they give very divergent results. On the whole, the so-called coefficient of similarity devised by Burt would seem to be the best for general use. However, such methods merely supply a general indication as to whether or not the factorial pattern exhibits any significant change, and should always be supplemented by a more detailed study of the material in order to take full account of all the complex issues involved. On applying one and the same battery to two populations at different levels, I myself found that, with Thurstone's procedure, the factorial composition of some of the tests appeared to change appreciably, and so modified the theoretical pattern, although for the majority of tests the saturations remained

fairly stable, and thus at first sight suggested that the factors extracted in the two investigations were in fact identical.

Nevertheless, even when no rigorous methods are applicable for identifying and comparing factors, the factorist should not confine himself merely to a superficial examination of the factor saturations and of the psychological content of the tests. He should also take into consideration the general conditions of the experiment as a whole. The question how far the influence of the general factor, *g*, diminishes with age, while at the same time special aptitudes become more and more plainly differentiated, provides one of the most typical examples of the difficulties arising in interpreting factorial results. Numerous researches have been carried out on this problem, and have led to very discrepant and even contradictory conclusions. In a brief analysis of these various inquiries, Vernon [22] has shown clearly that alterations in the factorial schemes depend not only on the composition of the test batteries, but also on the degree of heterogeneity or selection in the populations tested and on their educational or vocational training: he emphasizes the fact that, even if random samples of the population could be tested at different age levels, the content of the tests would also have to be adapted to the different ages, in order not to introduce modifications into the variance of the first factor. It would seem, therefore, that the influence of age presents a problem which is peculiarly difficult to solve.

The Nature of the Factors

When considering the problems raised when we attempt to determine the nature of the factors, it is well to bear in mind the view expressed by Thomson. "I do not believe", he writes, "that the factors arrived at by minimizing the number of common factors and maximizing the part played by specifics can be in any sense realities, either in a physiological or a psychological sense. I think that the causes of ability to do well in this, that, or the other test are innumerable and small, some of them genetic, some of them acquired, and that the correlations or resemblances between test-lists are due to some of these causes being helpful in both tests, which overlap in the sample they draw on from these innumerable little causes, which I would not dream of calling 'factors', and which are different in different persons. It remains true, however, that factors form a vivid and convenient description of any battery of tests, a description we would be silly to forgo, but also a description we must keep in its place and not raise to the status of a causal explanation" [23]. All this appears to me decidedly helpful, because it is likely to induce the psychologist to recognize the difficulties and the limitations, as well as the possibilities, of factorial methods.

Moreover, if factors are thought of as possessing essentially a descriptive role, then the divergences between the different factorial techniques will no longer seem irreconcilable. "The big difference", as Thomson himself points out, "between Thurstone's American school of thought and the view of a majority of British psychologists is that the latter, following the Spearman tradition, cling to a general factor *g*, whereas Thurstone's primary factors include no *g*.

The difference is not fundamental except to those who reify their factors and look upon them as actual entities, possibly with physiological bases such as a certain brain centre, a certain endocrine gland, or the like. Others, who look upon factors only as convenient descriptive parameters, realize that the two descriptions, with or without a *g*, are equally capable of accurately specifying a test or a man."

From a theoretical standpoint, the antagonism between the two schools is now tending to disappear. In adopting oblique factors, Thurstone has indirectly recognized the existence of a general factor; and he has since explicitly reintroduced such a factor in his recent attempt to re-analyse his centroid factors in terms of what he calls 'second order factors' [14]. Possibly the differences between these alternative types of analysis might have greater practical significance, were the factors themselves estimated in detail for each of the subjects tested. In practice, however, psychologists rarely make such estimates, and appear content to employ batteries containing groups of tests that are presumed to correspond with certain important factors, e.g. the verbal factor or the spatial factor; as a result, no matter what method of analysis may have been relied on at the outset when constructing such batteries, the groups of tests which they contain seem to all intents and purposes to be very much the same.

The view that the relations between tests are to be interpreted in terms of a large number of small causes suggests a reason why factor analysis does not help us to 'discover factors' and why it has proved so difficult in the course of such investigations to secure a small number of factors that are really stable. Beyond a doubt, no psychologist at the present day would pretend that it is sufficient to apply a large variety of tests to a large number of persons in the hope of determining, in a single research, all the 'factors of the mind'. Nevertheless, many investigators still embark on factorial inquiries without clearly realizing that the factors extracted will depend largely on the choice of the tests which they put into their batteries, and that factorial analysis cannot be safely employed except in strictly limited types of experiment, namely, those that have been deliberately designed with the needs of a factorial study in view. It is therefore only valid when, on the basis of some initial psychological hypothesis, a provisional classification of the data has been explicitly laid down at the outset. It is this constant neglect of the intrinsic limitations of factorial procedures that appears to be the main reason why so many arduous investigations now seem to have been entirely wasted.

Conclusions

The main conclusions that emerge from this broad review may be briefly summed up as follows. Factorial analysis is not a mode of psychological research which is sufficient in itself. Unaided, such a method, as Burt has said [24], does not permit a determination of causal influences: these can only be discovered "by means of experimentation, of controlled observation, and of introspection". Factor analysis is, in his view, essentially a statistical technique which allows us to verify and specify with greater precision a hypothetical classification, and which may, in certain cases, tend to modify or even to suggest such a classification.

It is in short a mathematical test like calculating Student's *t*, or chi-squared, or an analysis of variance. Confined to this role it provides the psychologist with a supplementary aid of great value which can be used *in the course of*, but not as the sole basis of, a psychological research. It is to be hoped that a further development of factorial techniques will render such verificatory tests more exact and more effective.

References

[1] Spearman, C. (1914). The theory of two factors. *Psychol. Rev.*, XXI, 101–115.

[2] Burt, C. (1909). Experimental tests of general intelligence. *Brit. J. Psychol.*, III, 94–177.

[3] Board of Education (1924). *Report on Psychological Tests of Educable Capacity*. London : H.M. Stationery Office.

[4] Spearman, C. (1927). *The Abilities of Man*. London : Macmillan.

[5] Delaporte, P. (1947). Prolongement de la méthode d'analyse factorielle de Spearman en utilisant la statistique mathématique. *Biotypologie*, IX, 45–59.

[6] Burt, C. (1949). The two-factor theory. *Brit. J. Psychol.* (*Stat. Sect.*), II, 151–179.

[7] Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis : a Synthesis of Factorial Methods*. Chicago : University of Chicago Press.

[8] Burt, C. (1950). Group factor analysis. *Brit. J. Psychol.* (*Stat. Sect.*), III, 40–75.

[9] Burt, C. (1949). Subdivided factors. *Brit. J. Psychol.* (*Stat. Sect.*), II, 41–63.

[10] Thurstone, L. L. (1931). Multiple factor analysis. *Psychol. Rev.*, XXXVIII, 406–427.

[11] Thurstone, L. L. (1935). *The Vectors of Mind*. Chicago : University of Chicago Press.

[12] Thurstone, L. L. (1951). L'analyse factorielle : méthode scientifique. *Année psychologique*, L, 61–75.

[13] Emmett, W. G. (1949). Factor analysis by Lawley's method of maximum likelihood. *Brit. J. Psychol.* (*Stat. Sect.*), II, 90–97.

[14] Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago : University of Chicago Press.

[15] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, II, 6 Ser., 559–572.

[16] Burt, C. (1949). Alternative methods of factor analysis and their relations to Pearson's method of ' principal axes '. *Brit. J. Psychol.* (*Stat. Sect.*), II, 98–121.

[17] Burt, C. (1952). Tests of significance in factor analysis. *Brit. J. Psychol.* (*Stat. Sect.*), V, 109–133.

[18] Bernyer, G. (1956). Analyses factorielles de deux batteries de tests appliquées aux candidats d'une école de mécanique. *Biotypologie*, XVII (in the press).

[19] French, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs*, No. 5. Chicago : University of Chicago Press.

[20] Barlow, J. A., and Burt, C. (1954). The identification of factors from different experiments. *Brit. J. Stat. Psychol.*, VII, 52–56.

[21] Leyden, T. (1953). The identification and invariance of factors. *Brit. J. Stat. Psychol.*, VI, 119.

[22] Vernon, P. E. (1950). *The Structure of Human Abilities*. London : Methuen.

[23] Thomson, G. H. (1951). Factor analysis, its hopes and dangers. *Proc. XIIIth Internat. Cong. Psychol.* Stockholm.

[24] Burt, C. (1947). L'analyse factorielle dans la psychologie anglaise. *Biotypologie*, IX, 7–44.

Vol. X Part I The British Journal of Statistical Psychology May 1957

THE COMPARISON OF FREQUENCIES IN MATCHED SAMPLES

By Alan Stuart
Research Techniques Unit, London School of Economics

Abstract

For purposes of statistical comparison, matched samples are in general more accurate than unmatched. In testing the statistical significance of the results, it is essential to take into account the gain in accuracy, due to the matching. The following paper indicates how this can be done in the case of attributes.

1. Introduction

Psychologists and others often seek to eliminate the effects of unwanted sources of variation on comparisons between two groups by 'matching' individuals, one from each group, for values of the variables to be eliminated. If a quantitative variable is being investigated, differences between matched individuals may be tested by a Student's t test or some analogous distribution-free procedure. A similar test is required when an attribute is being investigated, but does not seem to have been made available. It turns out that such a test can be very simply derived (almost entirely by reference to the ordinary 2×2 table), and that it yields a test statistic whose distribution is approximately normal. A numerical example will be given in the last section.

2. The Standard Test

Let us write the 2×2 table in the usual form as follows:

	Number possessing attribute	Number not possessing attribute	Sample size
First sample	a	b	$a+b$
Second sample	c	d	$c+d$
Combined samples	$a+c$	$b+d$	$a+b+c+d=n$

The standard test for the equality of the proportions in the two groups from which random samples were drawn is made by computing

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}. \qquad (1)$$

If all the cell frequencies are of reasonable size (say 5 or more), this will be distributed with a close approximation to the chi-squared distribution with 1 degree of freedom.

Another way of looking at the test is to remember that a chi-squared variate with 1 degree of freedom is the square of a standardized normal variate. Thus

$$\chi = (ad-bc)\sqrt{\left\{\frac{n}{(a+b)(a+c)(b+d)(c+d)}\right\}}$$

is approximately such a variate. Let us re-arrange (2) to give the algebraically identical equation:

$$\chi = \frac{\dfrac{a}{a+b} - \dfrac{c}{c+d}}{\sqrt{\left\{\dfrac{a+c}{n}\cdot\dfrac{b+d}{n}\cdot\left(\dfrac{1}{a+b}+\dfrac{1}{c+d}\right)\right\}}}. \tag{3}$$

We then see that χ is simply the ordinary test for the difference between the observed proportions $a/(a+b)$ and $c/(c+d)$, the difference being tested against its standard error estimated from the combined samples.

If the samples are of equal size, namely, $a+b = c+d = \frac{1}{2}n$, then (3) becomes

$$\chi = (a-c)\sqrt{\left\{\frac{n}{(a+c)(b+d)}\right\}}. \tag{4}$$

3. The Effect of Matching

We now examine the effect of a matching procedure upon our test statistic. As before, we wish to test the difference in the numerator of χ in (3). On the hypothesis that there is no true difference between the matched groups, that numerator will have a mean value of zero, as in the unmatched case; but its standard error will be reduced by the matching process, which has the effect of a stratification upon a random sample. In the denominator of (3) the factor

$$\frac{a+c}{n}\cdot\frac{b+d}{n} = \frac{a+c}{n}\left(1-\frac{a+c}{n}\right), \tag{5}$$

which is an estimate of the ordinary binomial (Bernoullian) variance, should therefore be replaced by the unbiased within-strata (Poissonian) estimate,

$$2n^{-1}\sum_{i=1}^{n/2} 2\{\tfrac{1}{2}(x_i+y_i)\}\{1-\tfrac{1}{2}(x_i+y_i)\}, \tag{6}$$

where x_i, y_i refer to the first and second samples respectively, and take the value 1 when the attribute is present in an individual and 0 when it is not. Equation (6) simplifies to

$$\frac{2}{n}\sum\{(x_i+y_i)-\tfrac{1}{2}(x_i-y_i)^2\}; \tag{7}$$

and this becomes

$$\frac{2}{n}\{(a+c)-\tfrac{1}{2}\sum(x_i^2+y_i^2+2x_iy_i)\}, \tag{8}$$

since $\sum x_i = a$ and $\sum y_i = c$. Moreover, since x_i and $y_i = 0$ or 1 for all i, we may further simplify (8) to

$$\frac{2}{n}\{(a+c)-\tfrac{1}{2}\sum(x_i+y_i+2x_iy_i)\} = \frac{1}{n}\{(a+c)-2S\}, \tag{9}$$

where we have written $S = \sum x_i y_i$. Here S is the number of matched pairs in which *both* individuals possess the attribute. Substituting (9) for (5) in (3), and putting $(a+b) = (c+d) = \frac{1}{2}n$, we obtain

$$Y = \frac{a-c}{\sqrt{(a+c-2S)}} \tag{10}$$

as our standardized normal variate for the matched case. In testing for significance Y is referred to the usual critical values of the normal distribution.

Unlike (4), its analogue in the unmatched case, (10) does not explicitly contain n, the effect of total sample size being expressed by the other terms. Moreover—though this is not at once evident from inspection—(10) is symmetric as between a and c on the one hand and b and d on the other. If we transpose the columns of the 2×2 table, we get exactly the same result, as we should for any reasonable test.

4. Sensitivity of the Test

Whether (10) provides a sensitive test in any particular situation will depend upon whether the matching procedure has been efficacious. Suppose, for instance, that the matching makes no difference to the joint occurrence of the attribute, i.e. suppose that the probability that an individual whose matched 'twin' possesses the attribute will himself also possess it, is exactly the same as the overall probability of possessing the attribute, namely, $(a+c)/n$: then the expected number of matched pairs, each possessing the attribute, will be

$$\mathscr{E}(S) = \tfrac{1}{2}n\{(a+c)/n\}^2 = (a+c)^2/2n. \tag{11}$$

If in (10) we substitute (11) for S, we obtain for Y precisely the value of χ in (4). That is just what we should expect.

Clearly, if S tends to be larger than the value given by (11), then (10) will be increased and the test made more sensitive. Conversely, if it tends to be smaller, we may actually lose efficiency by matching, since the individuals of a matched pair may behave more differently than if unmatched. This is not, of course, a contingency likely to arise very often in practice; but it is worth mentioning to underline the fact that the benefits of matching are not automatic. In normal circumstances, where it is reasonably certain that matching has done its work properly, it would consequently be reasonable to use (4) as an effective lower bound for (10) in cases where for some reason S cannot easily be computed. This will be less wasteful of information than simply ignoring the term $2S$ in (10).

The extreme example of the gain in sensitivity which may accrue as the result of matching will occur when S takes its maximum value, which will be the lesser of a or c. In this case, (10) gives

$$|Y| = \sqrt{|a-c|},$$

which is significant at the 5 per cent level ($\pm 1{\cdot}96$) whenever $|a-c| \geqslant 4$.

5. Correction for Continuity

In using the standard test statistic (1), we can improve the approximation of its distribution to the chi-squared distribution by applying Yates' correction for continuity. The effect of this is to replace the term $(ad-bc)^2$ in the numerator of (1) by $(|ad-bc|-\frac{1}{2}n)^2$. The same principle applies here; and in (10) the correction will have the effect of replacing $|a-c|$ by $|a-c|-\frac{1}{2}$.

6. Example

The following illustration is based on fictitious data. 50 boys and 50 girls, matched in pairs by age, size of family, father's educational level, and number of rooms per person in the home, were studied in order to determine whether there is any sex-difference in performance at the 11 plus examination. The results were as follows:

Pupils	Obtained Grammar School place	Did not obtain Grammar School place
Boys	5	45
Girls	13	37

The value of S, the number of matched pairs in which both boy and girl obtained a grammar school place, was 4. We have, from (10),

$$Y = \frac{5-13}{\sqrt{\{5+13-(2\times 4)\}}} = -2{\cdot}53.$$

Although the difference is small, it is fully significant at the 5 per cent level, and indeed almost at the 1 per cent level. The effect of the continuity correction would be to reduce Y to about $-2{\cdot}3$.

I am grateful to Mr. J. Durbin for suggesting several improvements in the presentation of the paper.

Vol. X Part I — The British Journal of Statistical Psychology — May 1957

HEREDITY AND INTELLIGENCE: A REPLY TO CRITICISMS

By Cyril Burt and Margaret Howard
University College, London

Summary

An attempt is made to remove current misconceptions about the way in which individual differences in general intelligence appear to be inherited. Further support for a multifactorial form of the Mendelian theory is found in an examination (*a*) of the commoner objections, both methodological and statistical, and (*b*) of the chief alternative hypotheses, viz., (i) the theory of blending inheritance and (ii) the theory of exclusively or predominantly environmental determination.

I. The Commoner Types of Criticism

Nature and Importance of Problem

In several earlier publications (Burt [2], [5] and refs.) one of us endeavoured to show how effectively the Mendelian principles of inheritance seemed to explain the observable distribution of mental differences, especially differences in what is commonly termed 'general intelligence'. During the past few years the problem of innate differences has come once again to the fore, particularly in connection with the allocation of pupils under the Education Act of 1944 and the Report of the Royal Commission on Population in 1949. In the meantime the psychological surveys carried out in London, Bath, Scotland, and elsewhere have furnished new data; while the recent advances in genetics, the fuller understanding of the mechanisms of heredity, and above all the improvements in biometric techniques, now make it possible to re-examine the basic issues with greater hope of success.

In the preceding issue of this Journal [9] we endeavoured to defend, on the basis of a more precise analysis, a theory previously put forward in general outline—namely, that innate differences in intelligence are due to Mendelian factors of two kinds: (i) major factors responsible for comparatively *large* deviations, usually of an abnormal kind, and (ii) multiple factors whose effects are *small*, *similar* and *cumulative*. However, the hypothesis that intelligence is inherited, both in the form that we had suggested and in various other versions, has of late been subjected to vigorous attack. In the main, the criticisms appear to fall into three more or less overlapping groups: first, those directed against the psychological and biological concepts in terms of which the hypothesis is expressed; secondly, those arising out of the abstract methodology of such inquiries; and thirdly, those concerned with the statistical techniques. As a rule, each critic singles out merely those particular aspects that happen to interest him, and ignores the other lines of argument. Nothing like a systematic survey of the various difficulties has so far been attempted. Some such review

seems urgently needed, since a study of the commoner objections and misunderstandings may help to set both the theory and the reasons for it in a sharper light, and may at the same time obviate a number of practical proposals which appear doomed to disappointment.

II. General Misconceptions

The Meaning of Heredity

The most frequent type of criticism springs from misunderstandings about the exact nature of the issues involved and of the hypothesis put forward. Two distinct questions are constantly confused : (i) in what way, if at all, are genetic differences in ability transmitted ? (ii) how accurately do existing tests or other methods of estimation enable us to assess such differences ?

In the first place, both critics and supporters of the doctrine of mental heredity seem as a rule to hold quite obsolete and misleading notions about the way biological characteristics are transmitted. Tacitly if not explicitly, ' heredity ' is assumed to mean ' the tendency of like to beget like ' (the definition quoted by one critic from the *Oxford English Dictionary*). Consequently, the arguments for inheritance are supposed to consist in demonstrating resemblances between the parent and his children. When the two parents differ, the child is supposed to consist in a blend of both—rather on the lines of what the Arabian sage declared would happen if you mated a white elephant with a black giraffe, or the views that led older biologists to predict such hybrids as the centaur, the mermaid, and the minotaur, and induced even Aristotle to describe the giraffe as the progeny of a camel and a leopard.

The approach of the modern geneticist is the reverse of all this. As he sees it, the real problem is rather to explain why in so many instances ' like begets unlike '. And the only satisfactory answer is that supplied by the 'mosaic theory of inheritance ' which we owe to Mendel :

> An atomic mechanism with unlimited power
> To vary the offspring in character, by mutual
> Inexhaustible interchange of the transmitted genes[1].

On this basis the chance re-combinations of a definite number of unalterable factors will yield, as a consequence of sexual reproduction, a wide variety of patterns in the ensuing generation, as dissimilar as the figures formed by shaking the coloured chips in a child's kaleidoscope.

As one of us has argued elsewhere, " probably the most convincing proof of the genetic determination of abilities is the appearance of extremely dull children in the families of the well-to-do professional classes and of extremely bright children in families where both the cultural and the economic conditions of the home would, on the environmentalist's hypothesis, condemn every child to manifest failure ". With the Mendelian hypothesis this is just what we should expect. An important corollary follows. On this hypothesis a far

[1] Robert Bridges, *The Testament of Beauty*, iii, 172–4.

greater proportion of the child's inherent characteristics are transmitted to him through his parents than we should infer simply from the resemblances between them as measured by coefficients of correlation.

There is a further reason why a preoccupation with mere resemblances may easily be misleading. The family is the chief instrument for transmitting, not only innate biological characteristics, but also those intellectual, moral, and social characteristics that are acquired afresh by each generation. A few of our critics expressly recognize this fact, and even turn it against us. But here their arguments merely echo what we ourselves have always stressed. It was a point emphasized in the very first of our contributions ([2], p. 172); and we regarded it as a conclusive reason for not accepting test-results at their face value. The older types of test were particularly open to such influences. In the earliest studies of the Binet tests, for instance, it was noted that " tests which depend on a wide vocabulary or on items of skill and information imparted during early life in a cultured home " are apt to favour the child of intellectual parents regardless of his innate capacity; on the other hand, certain tests of a performance type appeared, if anything, to favour the child from the non-intellectual type of home ([4], pp. 194f., and [6], p. 59).

It was largely for this reason that, in formulating our theory, we have always attached special weight to the distinction between genotype and phenotype—a distinction which practically all our critics ignore, although it is basic to the whole Mendelian theory. There are, in fact, no such things as hereditary *characters*; there are only hereditary capacities or tendencies : and the modern geneticist continually insists that " almost every character can, and in fact nearly always does, show heritable and non-heritable variation simultaneously " [12]. To define ' intelligence ' as ' what intelligence tests measure ', and then argue that " the family background and the environmental influences that result are more than sufficient to account for all the observed differences in a mental character like intelligence "—as Dr. Woolf, Dr. Heim, Dr. Fleming, and several other critics seem to do—is to miss the main point of our argument. No doubt, a few human characteristics, such as eye-colour or serological type, are virtually unaffected by environmental conditions; and these, as we have suggested, might profitably be used as ' markers ' to determine in fuller detail the transmission of mental tendencies. But no traits in which the psychologist is interested are likely to be determined either exclusively by inheritance or exclusively by environment; and for much the same reason the mere fact that a particular trait is demonstrably affected by environmental and post-natal conditions is no proof that it may not also be profoundly influenced by heredity.

The wide-spread notion that the evidence for mental heredity rests almost entirely on the study of family resemblances seems largely due to the work of Karl Pearson. Pearson, unlike Galton, strongly opposed what Galton had called the ' particulate theory ', and clung firmly to the traditional doctrine of ' blended inheritance '. On this hypothesis it was natural to infer that the coefficient of correlation, as an index of resemblance, supplied the best measure

of heredity. Even today most psychological and educational writers who discuss the subject still implicitly accept the Pearsonian conception; and few seem to realize that any other interpretation of heredity is available. As a result they commonly take for granted that anyone who employs Pearson's statistical procedures, and calculates coefficients of correlation, must necessarily subscribe to Pearson's hypotheses and conclusions. Dr. Fleming [16], for example, describes " Pearson, Burt, Terman, and other workers in the first few decades of the twentieth century " as " believing in the overriding importance of hereditary differences ", and apparently supposes that all of them relied on the same type of evidence and adopted the same preconceptions. Dr. Lewis [24], in criticizing the application of a multifactor theory to such traits as intelligence, similarly assumes that its advocates start with the postulates laid down by Pearson. Both critics have evidently overlooked the objections which Burt [2] urged against Pearson's sweeping generalizations, and the arguments which he adduced in favour of the Mendelian doctrine, which Pearson with equal emphasis repudiated.

The Need for Specific Investigations on Specific Characteristics

There is another cautionary principle to be learnt from Mendel. Pre-Mendelian discussions of heredity nearly always dealt with organisms as wholes. When a hybrid animal was bred from two different parental strains, earlier biologists (including Darwin himself) were content to note the general similarity between the different creatures, much as they did in deciding whether a newly discovered animal or plant should be assigned to this or that species. Mendel, on the other hand, in his work on hybridization, announced at the very outset that it is essential to deal separately with separate characters. Psychology has been slow to adopt the same principle. In their references to the genetic differences between the sexes, for example, psychologists, educationists, and sociologists, even in the earlier decades of the twentieth century, still argued the question in the most comprehensive terms : do women display an all-round inferiority to men, or are they the equals of the male ? The notion that the average woman might be genetically inferior in some capacities, superior in others, and perhaps qualitatively different in yet other respects, or that certain characteristics might be definitely sex-linked, has been confined chiefly to what Dr. Maberly calls the ' psychometric school ' [3]. Most medical psychologists, indeed, sharply contest the ' atomistic view ' on the ground that " each personality, whether male or female, must be viewed as a unique and unanalysable whole, not as a mosaic of particles or factors ".

A number of our critics, however, somewhat inconsistently assume that ' the hereditarian school ' shares this monistic attitude. The psychometrist, it is said, seeks to assign ' a single overall proportion ' which will hold good of the entire range of human activity; and his ' belief in the overriding importance of hereditary differences ' is depicted as applying to every form of behaviour in an equal degree. But this version of our views is quite unwarranted by

anything that we ourselves have said. Probably the notion that " Spearman, Burt, and their co-workers use the data of their tests to deduce a single all-embracing estimate for the relative importance of nature and nurture " springs from a misinterpretation of the phrase ' general intelligence '. Spearman, it is true, in his earlier correlational studies maintained that his ' hierarchical tables ' demonstrated the presence of a single general factor only—' the sole determinant of an individual's potential achievements in every walk of life ', and held that specific abilities, or what were described as such, were simply the effects of acquired knowledge, interest, or skill. We ourselves would wholly agree with those who reject any such simple assumption. Nevertheless, as numerous studies of backwardness indicate, an upper limit for each individual child is undoubtedly set by this innate factor of intelligence, though it is a limit which few children, bright or dull, ever actually attain. Moreover, both limits and possibilities depend partly on special abilities; and, like intelligence, these abilities are also to a large extent genetically determined. Consequently, what is actually attainable in the different subjects of the curriculum will be limited in very different degrees.

Dr. Heim and Dr. Harrison apparently imagine that, when we concluded that " about 75 per cent of the variance shown by tests of intelligence is attributable to innate or inheritable causes ", the figure was meant to specify " the proportional contribution of heredity to the individual's general educational and social efficiency " in *every* direction. To maintain any such wholesale conclusion would, we readily admit, be " to maintain what is at once meaningless and absurd " [21]. Indeed the absurdity becomes obvious as soon as we consider what such an inference would entail. Our figure of 75 per cent related to the variance contributed by the innate factor to the results of intelligence tests. We should therefore have to suppose that the ' saturation ' or ' loading ' for this innate factor in every school subject—reading, composition, arithmetic, manual work, and the rest—would not only be the same, but would always be unity. Now in various investigations on tests of educational attainments we have clearly shown that the saturation coefficient varies widely from one subject to another, and that it varies almost as widely for the same subject at different ages.

Spearman's view no doubt implied, as he himself pointed out, that, in theory, every individual could be ranked in a single series which would remain unchanged for all forms of cognitive activity. And as a result those who advocate the use of intelligence tests in the school have generally been credited with much the same doctrine. Not unnaturally, the notion that each human being is stamped at birth with his own serial number, and has his place irrevocably fixed in the queue for relative efficiency, is repugnant to most social writers of the present day, whose sympathies lie usually with those in the lower ranks. However, as must be clear from our previous publications, we should be the first to reject this modernized doctrine of predestination. Human beings, if they are to be classified at all, will have to be arranged, like the abilities they

exhibit, in a ramifying hierarchy, not in a one-dimensional scale : and if some particular individual is weak in one type of ability, there are others in which he may quite possibly excel. Thus our reply to the problem raised by Heim and Harrison would be that very succinctly expressed by Haldane : " the question of the relative importance of nature and nurture has no general answer, but only a very large number of particular answers " ([18], p. 36).

The Analysis of Assessments of Intelligence and the Analysis of Tests

However, in this paper, as in the preceding, we shall be concerned, not with special abilities, but only with general intelligence. Here once again it is to be remembered that what the psychologist can directly observe or measure is only the phenotype, never the genotype. Hence our first aim has been a practical one—to estimate the relative efficiency of the methods available for assessing intelligence. We might almost borrow Galileo's explanation when he repeated Archimedes' experiment : " I am not trying to discover how many ounces of gold there may be in the crown, but simply how much accuracy there is in the method ". For this purpose we have endeavoured to ascertain the comparative merits of the two most obvious procedures : first, the use of standardized tests, with the marks taken just as they stand; secondly, checking and correcting the marks so obtained in the light of the available evidence.

Now, what our critics chiefly wish to challenge is the efficacy of the tests in common use. Many are disturbed by the idea that, merely on the basis of a Binet I.Q., some unfortunate child may be permanently labelled ' dull ', ' subnormal ', or even ' mentally deficient '; others fear that grave injustice may be done to youngsters from the underprivileged classes, if selection for grammar schools is determined mainly by their performances in tests at the 11 plus examination. We ourselves, however, are perfectly willing to admit that, as a means of estimating genotypic differences, even the most carefully constructed tests are highly fallible instruments, and that their verdicts are far less trustworthy than the judgments of the pupil's own teachers; the outstanding merit of the tests is that, unlike the teacher's judgments, they are comparable from one school to another.

Accordingly, our investigations have been directed towards discovering, first, how far such tests are likely to be affected by irrelevant influences, and secondly how far existing methods of assessment can be improved upon by more comprehensive devices. Most psychological investigators have concentrated on the first of these two questions only. Almost inevitably the investigator who enters the schools from outside is obliged to confine himself to short and simple procedures—group tests of the usual type. And here lies the unique advantage of an officially appointed psychologist who is himself a member of the education authority's staff : he not only knows at first hand what are the conditions affecting the examinees in the various schools, but also possesses the authority to extend or repeat his tests and his interviews, and to require from teachers the detailed information and the further assistance that he needs.

The data we have used for the purpose of our various analyses have been secured in this way; and, having satisfied ourselves that by these means we can reduce the disturbing effects of environment to relatively slight proportions, we have gone on to inquire whether the data so obtained—the frequency curves, the bivariate distributions, the correlations between relatives brought up under various conditions (twins, cousins, children and their parents, siblings living in the same home, in different homes, in residential instititutions, and the like)—are consistent with the hypothesis of multifactor inheritance. That has formed our chief theoretical problem.

Intelligence and Social Class

Most of the critics we have cited appear to confuse these various issues. This is especially noticeable in discussions on the relations between intelligence and social class. The hereditarian holds that, in communities where there is a good deal of freedom to move from one socio-economic group to another, " the class of an individual depends largely, though by no means wholly, on his innate characteristics and on those of the family from which he springs " [27]. The environmentalist retorts that the class of an individual is " the main factor determining his performance in those tests of intelligence by which his innate characteristics are supposedly assessed ". After all, it is said, intelligence tests are for the most part designed by academic psychologists, that is, by persons with the cultured outlook of the ' privileged classes '. As a result, they have " a strong bookish bias, and are decidedly ' U ' " (i.e. upper class) " in their tone and vocabulary " ; and to assume that such tests " measure intelligence and that what intelligence tests measure is innate " is (so Dr. Heim declares) to be guilty, not only of ' verbal imprecision ' but also of ' personal bias ' [21]. Professor Zangwill has recently come to the defence of the test-constructor, and suggested that his bookish bias may be justified by the fact that our present civilization itself is built up on books : in the near future, he thinks, a technological age may lead the psychologist to prefer a more practical brand of test. But the psychologist selects his tests and his test-items, not in accordance with his subjective notions, but on the basis of an objective item analysis and of the correlations of his tests with the main underlying types of ' cognitive ' (directive) process—perception, association, coordination, abstraction, reasoning, noetic synthesis, and the like. Unfortunately in actual practice the most effective modes of testing cannot always be used; and the defects of the tests adopted for workaday purposes result usually from the unavoidable need to adopt something that entails no great expenditure of either time or money.

All this, however, Dr. Heim seems to overlook. To support her views she cites the figures published in the inquiry on ' Ability and Income ' by Burt [7], which, she says, were " acclaimed as proof of the supremacy of heredity and the rightness of the existing order of things . . . . For such people " (the psychometrists of the hereditarian school) " equality of opportunity consists

in affording individual children the same opportunities as their parents"; and thus the investigator mistakenly infers that "'the haves' inherit from 'the haves'" their apparent intelligence, whereas in point of fact all that they inherit are "more books in the home" and "the means to pay for better schooling". The figures recently collected by Mrs. Floud [17] and Mr. Flann Campbell [10] in regard to the social composition of the child population in secondary 'grammar' schools have been interpreted in much the same way.

Now we cannot help feeling that it is Dr. Heim who is really guilty of "verbal imprecision and personal bias". When the psychometrist states that this or that group test 'measures intelligence', he merely means that it yields a more probable estimate of each individual's intelligence than could be obtained for the same group with the same expenditure of time and money. Dr. Heim interprets the phrase as though it meant 'this test measures intelligence *and nothing else*'. None of those she criticizes have ever assumed that "what intelligence tests measure is innate". And even the most hasty perusal of the article that she quotes will show that Burt and his co-workers were far from claiming that the figures there given supported "the rightness of the existing order of things". Indeed, their final conclusion was that, among the non-fee-paying population, "about 40 per cent of those whose innate abilities are of university standard *fail* to reach a university, whereas an equal number from the fee-paying classes receive a university education to which their innate abilities alone would scarcely entitle them" (*loc. cit.*, p. 98). Nor is it true to suggest that "the psychologists who design and recommend such tests are themselves members of the privileged classes". In point of fact, as they have gratefully recognized, those who have been most active in this field would themselves have been unable to enjoy the benefits of a public school or a University had it not been for the scholarship system; to quote Godfrey Thomson, this fact was for him (as it was for several others) the main "incentive . . . in endeavouring to improve methods of selecting pupils" ([37], p. xv). A glance at the reports of the L.C.C. Psychologist will show that in actual fact such tests have served far more frequently to save the backward and the merely retarded from being certified as mentally deficient and to award the bright pupil from a poor home a free place in a 'secondary' or grammar school when his performance in the older type of scholarship examination would probably have excluded him.

But in any case much of this criticism is really beside the point. No competent psychometrist would nowadays rely on results collected from tests applied primarily for practical purposes, in order to solve his own theoretical problems. At best, in an inquiry into the extent or nature of mental inheritance, the figures supplied by official examinations can furnish nothing but confirmatory evidence. And the material which we ourselves have chiefly used was obtained from surveys planned with the theoretical issues specifically in view.

Some justification for the protests made by Dr. Heim and other environmentalists can no doubt be found in the exaggerated claims of the older eugenists :

vehement assertions on the one side inevitably provoke vehement dissent on the other. Pearson, for example, more than once declared, on decidedly tenuous evidence, that " the influence of environment is nowhere more than one-fifth of that of heredity, and quite possibly not one-tenth of it " [31]. *Sed latet dolus in generalibus*: and this sweeping generalization trapped him into practical corollaries which not unnaturally tended to discredit the whole movement: " a heredity-hunt ", as G. K. Chesterton observed, " would be a greater menace than a heresy-hunt " [11]. A few years later Dampier Whetham, one of the most active champions of eugenics, stoutly opposed all ' competitive examinations ' on biological principles ; " scholarships " (he warned his readers) " have their dangers when used to raise those who win them out of their natural class; they were good sociologists, as well as good divines, who taught the child to ' learn and labour truly, and do his duty in that state of life into which it has pleased God to call him ' " [40]. " Francis Galton himself ", Professor Hogben reminds us, " was a man with a strongly aristocratic bias ". And perhaps it was not surprising if " eugenics thus became identified with a system of ingenious excuses for combating the amelioration of working class conditions " [22]. Yet, when all is said, it seems a little unfair to saddle psychologists of the present day with the exaggerated views and social prejudices of the pioneers of fifty years ago. Scientific issues cannot be solved by turning them into a kind of erudite class-war—patricians *versus* plebs. After all, the basic error committed by the earlier hereditarians was much the same as those of their present day critics. They equated ' observable intelligence ' with ' inherited intelligence '—the phenotype with the genotype; and, if their theory of inheritance was grossly at fault, that at least was pardonable when genetics was still in its infancy.

III. Methodological Criticisms

General Criticisms

The main methodological argument may be briefly summed up as follows. What the psychometrist calls ' genotypic intelligence ' is avowedly an abstract and hypothetical ' factor '. Now, hypothetical abstractions, like ' factors ', cannot be empirically measured; even if they could, the ' factors ' of the statistical psychologist are admittedly non-existent figments; and in any case all endeavours to establish the existence of an innate factor of intelligence have so far completely failed; therefore, any further attempt to investigate or measure it is plainly doomed to failure; and the quest may as well be abandoned.

Let us take these allegations one by one.

The Alleged Impossibility of Measuring Hypothetical Quantities

No one denies that strictly speaking " an innate quality, like intelligence, from its very nature cannot itself be measured ". But in that case, says the critic, " there can be little point in discussing a hypothetical innate ability unless we can measure it ", particularly as the only method of verification available must itself depend on the assumption that the quantities involved are

to be measured. This difficulty was urged by several contributors to the Symposium arranged by Section J (Psychology) of the British Association at its Newcastle meeting; and more recently Dr. Maddox and two or three of our correspondents have maintained [26] that any quantity which cannot be directly measured or 'operationally defined' must be 'metaphysical', and that it is "a postulate of modern scientific methodology that all such unverifiable concepts should be discarded out of hand, together with the ether and similar hypothetical figments".

Now we doubt whether the majority of scientists of the present day would subscribe to the 'postulate' on which this argument is founded, although twenty years ago the proposal to dispense with all but directly verifiable concepts enjoyed for a time a widespread vogue. Contemporary physics, particularly quantum physics, is full of hypothetical quantities which at best can only be estimated with varying degrees of accuracy. Energy, to take the most familiar concept, cannot itself be directly measured. Like intelligence, it is, by definition, a 'capacity', namely, a 'capacity for work'. Consider the ordinary methods of determining the amount of energy supplied, say, by a Leclanché or Daniell's cell. As the student of physics is warned, the laboratory methods commonly employed are valid only on the assumption that the energy from the chemical changes is wholly converted into energy of current. But, strictly speaking, this is never the case; and various 'errors of measurement' have always to be allowed for. This is exactly parallel to the psychologist's problem. The 'capacities' with which he is concerned can never be directly measured, but only assessed. Nevertheless, provided he can devise some means of judging the probable margin of error, this in no way precludes him from using such assessments in his statistical proofs.

The Alleged Non-existence of Mental Factors

Many of our methodological critics would readily allow that certain hypothetical quantities may, in some sense or other, be measurable; but they insist that such quantities form a strictly limited group. Even if there is no direct observational evidence that the quantity in question *does* exist, yet (so they argue) that quantity must be a thing that *might* exist; and, when it is introduced as part of a working theory in science, there ought in addition to be some factual evidence to show that it probably *does* exist. Critics who pursue this line usually quote Thomson's warning about the dangers of 'reifying' abstract factors: "the psychologist's factors are statistical artefacts; a man does not have a *g* in the sense that he has a liver" [37]. Moreover, as they point out, the variable that appears in our tables is "an abstract factor twice removed": it is the genetic component (or sub-component) of the general cognitive component. Such a rarefied abstraction, they contend, is "nothing but the ghost of a ghost", and can have "no real existence".

Here it seems helpful to distinguish between the mathematical model on which our proofs are based and the 'real' or concrete phenomena to which the

model refers. Practically all scientific theories are founded on symbolic models of this kind—highly simplified constructs formed out of highly abstract concepts, skeletons with the flesh all scrupulously removed. They are essential for the preliminary deductions; and, in order to apply or verify the corollaries so deduced, all that is logically necessary is that the algebraic variables defined in the model should correspond with, or, as the phase goes, ' be isomorphic with ', the essential features of the real phenomena. There is no need for them to be directly correlated with what is actually observed or measured: they may simply represent what Broad has called ' dispositional properties ' ([1], pp. 434f.)[1]. And, as he reminds us, it is distinctive of modern, in contrast to mediaeval science, to interpret causal properties of this kind (including what mediaeval writers would have called the ' powers ', ' faculties ', or ' capacities ' of the mind) in terms of the minute spatio-temporal structure of the material substances that presumably serve as vehicles for such properties. That, as we have argued elsewhere, would be a perfectly feasible way of interpreting the particular ' dispositional property ' which we have designated ' intelligence '.

But, of course, this version of the situation demands a separate set of proofs, to show how the factors in the model are related to, or represent, the characteristics in the actual persons that we test. Oblivious of this gap in their arguments, statistical psychologists are too apt to confine themselves to purely statistical considerations; and are often content to identify their ' factors ' straightaway with supposed psychological characteristics on no other evidence than that afforded by the names of the tests employed. In our view, such identifications always call for the support of non-statistical evidence—observational, introspective, experimental, neurological, or the like [8].

In an instructive paper, Mr. Miles has argued that hypothetical entities like the factor of intelligence may be described as having ' real existence ' " only when they are not ' factors of the mind ', but ' factors of the body ', i.e. dependent on the ' discoverable behaviour of genes or neurones ' " [29]. We fully agree that this is by far the most satisfying way of conceiving the dispositional properties that concern us here; and we hold that a plausible interpretation of this kind is actually furnished by the multifactor theory. Could this theory be adequately verified, we might even define an individual's innate intelligence (gamma) as proportional to the number of gamma-genes in his genetic constitution, where a gamma-gene is conceived as part of the ultramicroscopic structure of the chromosome, and may itself be defined as an

[1] For our purposes it will be sufficient to define a ' dispositional property ' as one which requires a hypothetical proposition to describe it. Thus, when we say that a steel spring has the property of ' elasticity ', we mean: " if such a spring is pressed, pulled, or strained, and then released, it will (unlike a piece of putty) return more or less to its original shape ". The phrase ' more or less ' implies that the property has varying degrees; and therefore, although by definition a hypothetical abstraction, is susceptible of measurement; in fact, the degree of elasticity may be expressed by various ratios or ' moduli ': e.g.

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\text{pull per area of cross section}}{\text{change in length per unit length}}.$$

allelomorph whose presence adds a small positive quantity to one and the same specifiable mental capacity, namely, the capacity to perform all kinds of cognitive task. It will no doubt be objected that to count the actual number of such genes would be an impossible feat. True; but it is equally impracticable to count the number of molecules in one cubic millimetre of a given substance; yet the molecular density of a gas is defined in terms of the number of molecules contained in such a volume. In both cases what the scientist does is to make, not a direct count, but an indirect estimate of the number in question.

The Alleged Non-existence of ' Innate Intelligence '

Dr. Maddox [26] endeavours to reinforce the criticisms we have been discussing by giving the methodological argument a more empirical turn. After prolonged research by various methods, he says, " investigators have failed to produce any convincing evidence about the genetic components of intelligence ". Hence our own "attempt to resuscitate the nature–nurture controversy " he condemns as quite " anachronistic ". A wiser methodology would leave such purely conjectural factors in the nebulous limbo to which they belong, and would concentrate exclusively on environmental factors. " Environmental factors ", he points out, " are capable of control; and for this reason are of greater interest to educators ". He therefore fully applauds the conclusion expressed by H. E. Jones " that potential contributors should now be advised that the nature–nurture controversy has been shown to be an unproductive field of research ".

This would be an easy way of cutting the Gordian knot. But a policy of contented nescience would mean renouncing every hope of solving numerous problems which are of first importance alike for theory and for practice. The mere fact that environmental conditions are more amenable to control, and hence of paramount interest to the educator, cannot of itself absolve the psychologist from the task of investigating the possible influence of genetic factors. In dealing with any particular child[1], it is as necessary to consider what cannot be done as to know what can. If Tommy's backwardness is due to an innate lack of intelligence, then it is useless, and worse than useless, to strive to transform him into a normal pupil. If Harry is not endowed with the ability requisite for the work in a grammar school, it is folly to transfer him thither simply because there is a slender chance that " even the dullest youngster may be a genius in disguise ". It is only in fiction that the hundred-to-one outsiders romp home in the final lap, and leave the experts gaping.

Our own explanation for the failure of previous research would be very different from that which Dr. Maddox has offered. The obvious reason surely is that the investigators to whom he refers were all thinking in terms of out-of-date concepts of heredity : even had they wished to substitute a Mendelian

[1] For conflicting opinions on this point, see the articles and correspondence on ' Dullness and its Causes ' in the recent issues of the Times Educational Supplement, October 19, 26, November 3, 10, 1956.

theory, they possessed no adequate techniques for coping with multifactorial problems. Dr. Maddox, however, apparently believes that, from the very nature of the issues involved, it is impossible to secure evidence that might justify a hereditarian interpretation ; and his scepticism, as he points out, can claim the sanction of most American psychologists. Nevertheless, it no longer commands the support of American geneticists. Professors Goldschmidt, Snyder, Calvin Hall, and Calvin Stone all deplore the neglect of genetic investigation by psychologists of the present day. Goldschmidt himself traces this general lack of interest to the fact that empirical psychology began as an offshoot of philosophy, at a time when associationism held the field : translated into physiological terms, the associationist theory gave rise to a ' doctrine of universal conditioning ', which has lately dominated the American behaviourist school; and as a result practically all individual differences are interpreted in terms of postnatal experience. " This ", he writes, " may have been responsible for the extreme belief of our contemporary psychologists in the power of environment ". Nevertheless, " with a little prodding from genetics, both human and animal psychology might increasingly assimilate genetic notions, and be stimulated by them for mutual benefit "[1].

What, then, in the most general terms, are the methods of the present-day geneticist, and what grounds have we for supposing that they cannot be profitably borrowed by the psychologist ? Broadly speaking, there are four main lines of approach : (i) cytological analysis—the study of cell structure ; (ii) the method of uniform environment with varied heredity; (iii) the method of uniform heredity with varied environment; (iv) genetic analysis—the study of the effects of the segregation and recombination of genes in successive generations (cf. [36]). Man is a slow-breeding animal of low fertility ; he cannot be readily subjected to experimental control; and his behaviour is governed by post-natal experience to a larger extent than that of any other creature. These peculiarities render the psychogeneticist's task more complex; but they do not make it impossible. The validity of such methods for the study of mental characteristics is a matter to be decided, not by the abstract canons of some preconceived methodology, but by the pragmatic principle of giving them a trial, and noting whether they yield any worth-while results.

The evidence we have offered in support of our own hypothesis has been derived from all four lines of approach. (i) In the case of man, cytological knowledge is still lamentably defective; yet it is sufficient to demonstrate that the mechanisms of heredity and variation are essentially the same in the human species as in all other bisexual creatures. We have relied on it chiefly to show

[1] *Genetics in the Twentieth Century* (1951), p. 7. In the same volume Professor Snyder recalls how, " after a short flurry of interest in human heredity during the first decade or two of the century, that interest, particularly in America, proved to be short lived " (*loc. cit.*, p. 370). Quite recently, however, centres of research in human genetics have been established in the United States, and by 1948 an American Society of Human Genetics had been formed. Although hitherto engrossed with the problems of physical and pathological heredity, its members are " already beginning to ask whether it may not be time to embark on genetic studies of normal mental characteristics ".

(*a*) that there is a strong antecedent probability that a basic character like intelligence, which appears to be continuously graded and normally distributed, is determined, or at any rate influenced by, a multiplicity of genetic factors and (*b*) that the number of genes or ' factors ' is large enough to provide the wide range of differences and the smooth distribution curves that are actually observed. (ii) Studies of children brought up in residential institutions where environmental conditions are relatively uniform have disclosed wide individual differences in intelligence that cannot possibly be ascribed wholly to environmental influences, and are closely correlated with those exhibited by the parents. (iii) Studies of monozygotic twins reared in different environments have been used to assess the effects of environmental differences on (*a*) intelligence tests of the ordinary type and (*b*) adjusted assessments based on all the information available for each individual child. (iv) In the main, however, the procedure adopted by the human geneticist must necessarily be statistical; and here the methods of genetic analysis developed by the modern biometric school have already proved their efficiency in analogous fields of work. Today, geneticists who work with cattle or with crops are constantly attempting to estimate the strength of whatever genetic quality it may be that makes for a good yield in this or that breed and in this or that strain : there can surely be no *a priori* objections to extending their methods, with appropriate modifications, to the problems of human psychology.

IV. Statistical Techniques

General Nature of the Techniques Employed

This brings us to the type of criticism most frequently advanced, namely, that directed against the biometric techniques that we ourselves have applied. During the earlier part of the century the statistical treatment of genetic problems, initiated by Galton, Pearson, and Weldon, was quickly overshadowed by the more direct and fruitful techniques introduced by the experimental and cytological investigators of the Mendelian school—particularly Bateson and his followers in this country and Morgan and his fellow workers in the United States. The interest in statistical research began to languish; and its neglect has been widely interpreted by psychologists and others as a sign of sterility. Today, however, practically all geneticists are fully agreed that, in attacking problems of multifactorial inheritance, statistical techniques are indispensable; and those developed by Fisher and his followers now command almost universal assent.

Yet they have not escaped all criticism, even at the hands of biologists themselves. In this country the most prominent critics have been Hogben and Woolf, who have paid special attention to the study of human intelligence. For an authoritative examination of their views the reader must be referred to the replies of the writers they have criticized. But a brief glance at the difficulties which they emphasize seems necessary here.

A. *Objections to the Analysis of Correlated Variables.* Several of our critics cite the final chapter of Hogben's book on *Nature and Nurture* [23] as if it completely disposed of any attempt to apply Fisher's methods to such problems as our own. Dr. Maddox, for example, in discussing what he describes as our ' balance sheet for nature and nurture ', claims that Hogben's arguments conclusively demonstrate that the whole procedure is ' logically untenable '. Hogben himself never goes so far as this. Indeed, in the chapter quoted, he expressly admits that " the technique of correlation can be used to draw attention to the existence of genetic differences or of differences due to environment ", and declares that " the importance of Fisher's hypothesis of cumulative factors is beyond dispute ". His chief contention is that Fisher's method of analysis, at any rate in its original form, is legitimate only when we are able to " deal with nature and nurture as independent variables "; when nature (or ' heredity ') and nurture (or ' environment ') are themselves correlated, then (so he maintains) the results may prove deceptive. " The data on which Fisher's own analysis for stature is based were ", he points out, " collected from a comparatively homogeneous social group ", i.e. from a group in which the relevant environmental conditions were much the same for all; but were we to apply it to " human populations in general ", we should be forced to recognize that the wide range of environmental differences might falsify any values " theoretically deduced on the assumption that the environment is homogeneous ". Hence, although such methods may safely be trusted to indicate the *presence* of genetic factors, we must be cautious about using them to indicate its *amount*; and thus, whatever significance Fisher's attempt to draw up " a balance sheet of nature and nurture " may have for the study of human characteristics, " it cannot entitle us to set limits to the changes that might be brought about by regulating the environment "[1].

In his larger volume on *Genetic Principles* Hogben sounds a more optimistic note. He assures us that " from a theoretical standpoint the prospect is brighter " than it formerly appeared. In his opinion, it is now both possible and necessary " to develop methods which will enable us to assess the relative importance of nature and nurture in a specified range of conditions ". Among the procedures available he includes the study of " observed differences in performance of psychological tests ", such as he himself has used in his investigations on twins; and he advocates five particular types of inquiry " which, when applied with proper safeguards and interpreted with discrimination, will ", he believes, " yield unequivocal data ". These are the study (i) of twins and siblings, (ii) of adopted children, and of the effects of (iii) inbreeding,

[1] The book quoted by Dr. Maddox and Dr. Heim consisted of semi-popular lectures, addressed to a medical audience nearly twenty-five years ago, and was concerned, not so much with the technical merits of Fisher's mathematical procedure, but rather with the practical inferences which had rashly been drawn from it. Hogben's own views are more systematically presented in the larger textbook cited in the next paragraph.

It should be added that, since the above was written, Dr. Maddox has slightly revised the wording of his argument: but his essential conclusions remain the same. Our quotations were taken (with his permission) from the earlier version.

(iv) outbreeding, and (v) order of birth; and under each heading he cites one or two model researches, usually from the domain of animal genetics. These lines of inquiry, it will be seen, are much the same as we ourselves had tentatively employed in our earlier investigations. And, though Hogben himself might not approve the details of our work, there is no difference of principle between us. In short, the prophet whom Dr. Maddox has called in to curse such enterprises seems more disposed to grant them his blessing.

Hogben ends with a plea which we ourselves should like to second as strongly as we can. After deploring the lack of " University machinery " for giving instruction in modern genetics to students of medicine, social psychology, and allied subjects, he goes on to urge, in the most convincing terms, the importance of independent provision and special centres for research in the field of human genetics. It is, as he says, a subject which demands " organized teamwork on a very large scale ", including the " collaboration of geneticists, statisticians, educational psychologists ", and other qualified persons. It will therefore be costly. Yet " what we know already liberally encourages us to hope we may eventually know much more ".

All this seems far more reassuring in its tone than Dr. Maddox's quotations might suggest. No psychologist would quarrel with the cautionary remarks which Hogben makes in criticizing the unwarranted deductions drawn by the older enthusiasts. We ourselves have never sought to apply the method to " human populations in general ". And the difficulties that arise when environment is allowed to vary unquestionably call for special treatment. They can, we believe, be satisfactorily met in two alternative ways: (i) by modifying the design of the experiment; (ii) by amplifying the statistical technique.

(i) When planning the inquiry, we may seek to exclude any irrelevant correlation with environment by selecting either the sample of persons or the sample of tests. Thus, when using tests that are not ' culture-free ', we could for purposes of theoretical research select children coming from a restricted type of environment in the hope of eliminating in advance any cultural differences that might affect the tests. On the whole, however, we think it best to substitute, where possible, tests (or other methods) of such a nature that cultural differences are not likely to affect the results.

(ii) In studying, for more practical purposes, the value of marks obtained with tests of the ordinary type, we have endeavoured to supplement Fisher's original procedure so as to cope with the further variables involved. To check the adequacy of the method, we proposed an alternative analysis in terms of the so-called parallelogram law. Here, instead of starting with the correlation between assessments and environmental differences, we deduced it; and we showed that in each case the coefficient so obtained agreed, as closely as could be expected, with the empirical value computed directly from ad hoc ratings.

Dr. Maddox himself raises one or two further objections to Fisher's procedure. Since, he says, when judged by the correlation actually obtained,

ancestral heredity will account for little more than 50 per cent of the total variance in bodily characteristics, Fisher is obliged to " invoke other factors, such as segregation and assortative mating, to avoid attributing the balance to environmental factors ", and as a result has grossly underestimated the effects of favourable and unfavourable environment and the probable interaction between the environmental and the genetic components. Moreover, so Dr. Maddox contends, even the " initial step "—namely, ascribing the correlation between children and their parents to genetic factors—" is suspect, because of the correlation of heredity with environment ".

To these further arguments there are three simple replies. (i) Segregation and assortative mating were not arbitrarily invoked. Assortative mating is an ascertained fact : and without segregation the hypothesis would not be Mendelian, but pre-Mendelian. Does Dr. Maddox really hold that Fisher should have discarded the concept of ' segregating factors ' in favour of the older theory of ' blending factors '—a theory rejected by every modern geneticist ? (ii) The data for bodily characteristics which Fisher used were obtained, as we have seen, from a " fairly homogeneous middle class group ". Hence with this particular group the supposed differences between " favourable and unfavourable environment " did not in fact exist. Further, it is quite erroneous to assume that the effect of environment inevitably increases the correlation found. As Yule points out, both environment and dominance may tend to reduce it. (iii) Finally, Fisher readily admits that some of the genetic components are only " imperfectly additive ", and has therefore introduced a specific procedure which enables him to treat the irregular effects of interaction as an independent residual component.

B. *Objections to Second Degree Statistics.* The most direct attack on the type of approach that we have used is contained in a provocative paper contributed by Dr. Bernard Woolf to the Edinburgh Colloquium on ' Quantitative Inheritance ' ([33], pp. 81–102). His argument is divided into two parts : in the first he challenges the principles and assumptions adopted by nearly every British writer who has discussed quantitative inheritance in man—Fisher, Fraser Roberts, Thomson, and Burt; in the second he examines the specialized techniques initiated by Fisher, developed by Mather, and " common to current workers in biometrical genetics ". The final outcome, so Dr. Woolf tells us, is to enforce two main contentions : " the first is that real genetic situations with environmental and other interactions include too many variables to be amenable to the kind of analysis developed by Fisher and his school; the second is the utter inadequacy of orthodox analysis of variance, when confronted with a real world not arranged in Latin squares with all effects strictly additive ".

(i) *The Inheritance of Intelligence.* All too often our various critics seem to take their notions of what their opponents have said, not from their opponents' publications, but rather from each other; and Dr. Woolf's somewhat garbled version of our views has been so readily accepted by later writers that it is necessary at the very outset to emphasize how often he misinterprets what we

have actually said. He opens with a detailed discussion of an ever-popular problem—the alleged decline of national intelligence; and attacks what he takes to be the solution propounded by those who adopt the ' multiple or additive factor hypothesis '. Nowhere does he explicitly state how he himself understands the term ' intelligence '; but his comments imply that he sides with " those who prefer to re-define intelligence as ' what intelligence tests measure ' ". However, little or no interest can attach to the question whether the mere capacity to answer tests has improved or diminished. The issue raised by the writers he quotes, and revived by the inquiries of the Royal Commission on Population, related to what the Commission itself called ' *innate* intelligence '. Burt's abridged memorandum, to which Dr. Woolf chiefly refers, gives a formal definition in the opening paragraph; and the Appendix ' On the Meaning of Intelligence ' emphatically rejects the alternative, which Dr. Woolf seems to prefer ([41], pp. 68–75). Of course, if Dr. Woolf insists on starting off with a ' re-definition ' of the crucial term, it becomes the easiest thing in the world to show that our argument leads to a *reductio ad absurdum.*

However, besides misinterpreting the key-word in the discussion, he also gives an unfair twist both to the evidence adduced and to the final conclusion. In summarizing the evidence on which his opponents chiefly relied, he tells the reader that " all their predictions have been *purely inferential*, based on *test scores* taken at *only one point* in time " (*loc. cit.*, p. 84, our italics). This statement contains at least three misleading implications. First of all, in the London investigations at any rate, ' the predictions ' were not based solely on ' test scores '. In the memorandum and elsewhere a clear distinction was drawn between ' the crude test results ' and the ' adjusted ' or ' final ' assessments. In doubtful cases different types of test were used at different ' points of time '; and in every case the assessments were submitted to the teachers, and, where necessary, corrected in the light of fuller information.

Nor is it true that ' the predictions ' were ' purely inferential '. If Dr. Woolf will turn to the section in the Memorandum entitled ' Estimating the Amount of Apparent Decline ', he will see that *two* methods of prediction are discussed. The first, headed ' Indirect Estimation ' (p. 69), *was* ' purely inferential '; but the calculations were immediately qualified by the comment that on such matters armchair deductions cannot be trusted without confirmation. Had the calculated changes actually occurred, they would have involved a marked increase in the number of defectives and an equally marked decrease in the number of scholarship winners; and this, it was argued, " the older head teachers and officials would assuredly have noticed ". An alternative method was therefore used in the hope of " securing direct confirmation ". This is described under the heading ' Direct Estimation ' (pp. 61f.), and consisted of repeated surveys carried out at two or more widely separated ' points of time '. The result was a very different set of figures. The conclusion ultimately reached was that " far more extensive studies are required ".

After sounding his ' stern warning ' against the ' fallacy ' of indirect inference, Dr. Woolf goes on to say : " Still less was it legitimate for Thomson and Burt, *without any tests whatever of the parents*, to visit upon them the marks scored by their children . . . . Their anticipations of decline might have been justifiable as pessimistic speculations, though the *great confidence* placed in them would not have been " (p. 86, our italics). But in each of our surveys, assessments were individually obtained for a representative sample of parents, checked, for purposes of standardization, by tests of the usual type ([41], p. 52, [8], p. 172). How else does Dr. Woolf suppose the correlations between children and parents were computed ? Moreover, the burden of the whole memorandum was to deprecate any feeling of ' confidence '. And the opening summary plainly announced that the evidence at present available was " far from conclusive ". Hence the reiterated plea for research.

In discussing the influence of genetic and non-genetic factors on individual differences in intelligence (or rather on the assessments of those differences), the " first conclusion " that Dr. Woolf wishes to enforce is that " Burt seriously underestimates the amount of population variance attributable to environmental influences ". To establish this point, be begins by examining " the study by Barbara Burks on which he (Burt) bases such far-reaching conclusions " (*loc. cit.*, p. 87). Dr. Burks' investigation, we are told, was carried out on only " 214 foster children with 105 children as a control group ". But the principal basis for Burt's conclusions was not this American inquiry, but a series of investigations conducted with the help of several co-workers, and covering nearly 600 children in " residential schools, orphanages, and foster homes where the inmates are received during early infancy " and including a far larger sample of controls attending ordinary L.C.C. schools. This itself was merely supplementary to a mass of other data—correlations between parents and offspring, siblings, twins, and the like—to which Dr. Woolf nowhere alludes. The American inquiry was only cited at the end of the Appendix to show how an independent investigator in an entirely different country had arrived at much the same figure as that derived from the London surveys.

Dr. Woolf next turns to Burt's table for the mean assessments for children classified according to their father's occupation. He admits that the total range, from those belonging to " higher professional or administrative " classes at one end of the scale to " casual labourers " and " institutional defectives " at the other, is " enormous . . . amounting to more than 30 points ". But, he asks, is it not possible to account for " all the observed differences by environmental effects . . . without any need to bring in genetical differences " ? To support this suggestion he quotes figures from various sources (including Burt and Thomson); but all that these figures show is that environment may appreciably affect uncorrected test-results—a point which of course was never in dispute. He " freely acknowledges that all these explanations " (i.e. explanations he himself would favour)—" differences in parental attention, birth-rank effect, pre-natal disadvantages, etc.—are speculative "; and certainly the data

he cites would seem rather to indicate that these various ' non-genetic influences ', though undoubtedly present, are far smaller than the genetic.

However, this is not our main objection. What we demur to is the tendentious way he frames his question. The implication is that there are only two alternatives : (i) that *all* the differences are to be accounted for by heredity; (ii) that *all* of them are to be accounted for by environment—a most misleading antithesis. Surely Dr. Woolf would not pretend that intelligence differs from every other human characteristic in being totally unaffected by heredity. And certainly he does not deny that differences in the apparent intelligence of children are associated with differences in the genetic constitution of their parents. Why then should we not invert his question, and ask : " is it not possible to account for all the observed differences by the genetical differences without bringing in environmental influences as well ? ".

But, of course, both questions fallaciously imply that the two modes of causation must be mutually exclusive. For such an assumption there is not the slightest ground. Indeed, it must be obvious to anyone, who is not already strongly biased towards one explanation or the other, that *both* types of influence are generally at work. And the question to be asked is not : *could* the difference be accounted for wholly in this way or wholly in that, but what is *the most probable* way of apportioning the results between the first type of cause, the second type of cause, and the mutual interaction of the two.

Throughout this part of his paper much of the apparent plausibility of the arguments employed rests on a fallacy so often committed by psychologists and others who have not themselves had experience of mental testing. He treats ' the I.Q. ' as a kind of absolute measurement which is necessarily the same for the same child regardless of the method by which it has been obtained. Thus conclusions drawn by one writer who is referring to I.Q.'s obtained with group tests are applied to other I.Q.'s obtained by another writer who used individual tests; conclusions drawn from verbal tests are applied to I.Q.'s obtained with non-verbal tests ; and conclusions drawn from uncorrected I.Q.'s obtained from any test are assumed to hold good of composite assessments, carefully checked, where the I.Q. forms merely a conventional unit. Hardly ever, when quoting an I.Q., does he mention how it was secured. With the original Binet tests (with which some of the quoted results were obtained) the standard deviation was less than 13 I.Q. points; with the tests used by Cattell it was over 20; for our adjusted assessments (based originally on percentile ranks) we used an arbitrary I.Q. unit with a standard deviation of 15. Dr. Woolf treats these and other I.Q.'s as though they were all equivalent. That, of course, would make nonsense of any re-allocation of the variances.

(ii) *The Analysis of Variance.* In attempting to assess the relative influence of genetic and non-genetic factors, we ourselves relied almost entirely on an analysis of variance; and the second part of Dr. Woolf's paper is devoted to a spirited attack on this procedure.

His first and main objection to what he describes as ' the red herring of variance analysis ' is that it commits us almost inevitably to an underestimate

of those environmental influences that he himself has enumerated. It is, he says, a necessary condition of the statistical technique adopted by Fisher and his school that " environment should appear as a random error which can then be eliminated by arithmetical legerdemain rather than as an integral part of the situation to be embodied in the conclusions ". Here, beneath the phrase ' rather than ' there lurks the same fallacy as before—the tacit substitution of the ' exclusive (or alternative) ' *or* for the ' inclusive '. From the standpoint of the statistician, environment may be conceived as influencing individual variations in two distinct ways—first, by its ' systematic ' effects (as ' an integral part of the situation '), and secondly by its ' random ' effects. These are two *separable factors* (though in fact they are usually found together); they are not two *alternative methods* of treating one and the same factor. Because they are separable, it is possible, by a suitable choice of cases, to get rid of the systematic effects before ever we proceed to the statistical analysis. When this has been done, it is no longer true to say that what is an ' integral part of the situation ' now ' appears ' as a random factor. It has been eliminated in advance; and the ' random ' part is merely what is left. Moreover, in the material with which Fisher was dealing, the elimination could not have been the result of Fisher's statistical technique, since it had already been effected by the manner in which Karl Pearson had actually chosen his sample.

In another part of his paper, Dr. Woolf seems vaguely aware of this point, and puts his criticism in a different form. If, he says, we intend to assume (as Fisher does in the case of bodily measurements and as we have done in the case of intelligence) that the trait to be analysed is " influenced by *both* heredity *and* environment, then environment *must* be given primary place in the analysis : it is not legitimate to postulate genetic differences until it has been shown that environmental influences are unable fully to account for the variations in the trait; even then, heredity can only be invoked to account for the residual differences " (our italics).

Now, had the researches which Dr. Woolf is criticizing been designed to study the various processes by which environment can affect the phenotypes, then the principle he thus prescribes would no doubt have been appropriate. But their purpose was quite different. Fisher's object was to demonstrate that his own theory of human inheritance accounted for the data obtained by Pearson far better than the theory that Pearson himself had championed. We in turn were anxious to see whether much the same form of the multifactor theory might not account for the parallel figures obtained with assessments of intelligence. But we had an ulterior aim as well, namely, to determine, within rough limits, what proportion of the total variance was attributable to genetic factors in the case of (i) raw measurements obtained from tests and (ii) adjusted assessments based on all the evidence accessible for each individual child. Surely it is sound methodology to eliminate so far as possible, and for the rest to randomize, environmental effects when the object is to analyse heredity, just as it would be to eliminate or randomize genetic effects if the object was to analyse environment.

Dr. Woolf, however, declares that " randomization of environmental effects is tantamount to saying that there is no hope of ever understanding that aspect of the situation ". Nevertheless, in other researches—e.g. Burt's studies of backward children and of those mistakenly diagnosed as mentally deficient, where the problem was to analyse the manifold ways in which environment might mould the child's mental development and modify his performances in tests of various types—the operation of ' that aspect of the situation ' *was* very intelligibly explained. In any case, this part of Dr. Woolf's argument surely flies in the face of his other contention—the contention which he himself places first. He started, it will be remembered, by declaring that " real genetic situations . . . contain too many variables to be amenable to the kind of analysis developed by Fisher and his school ". And that is precisely why, on our view, when planning such inquiries, it is essential to simplify the issues, and deal with separate aspects of the situation in separate inquiries. But, if in the same research we try to embody environment ' as an integral part of the conclusion ', that must inevitably multiply the number of variables instead of diminishing it.

In order to demonstrate in greater detail the fallacies involved in using ' second order statistics ' (i.e., variances and coefficients of correlation) Dr. Woolf proceeds to scrutinize at some length the illustrative example worked out by Mather in his book on *Biometrical Genetics*. This he takes to be typical of the methods adopted by current workers in that subject. He himself contends that in all such investigations ' first order statistics are essential '. Why precisely first order statistics should be exempt from the disadvantages of second order statistics is left rather in the dark. Indeed, as Dr. Mather observes, Dr. Woolf's mode of argument seems to " contribute more to a trenchant style than to a scientific appraisal "; and in his criticisms he too often " actively misrepresents both the assumptions made and the methods used ". In regard to the issues with which we ourselves are concerned, there would seem to be little difference between the results obtained by the two types of statistic. Suppose, for example, we wanted to determine whether the test-performances of children reared in foster-homes were superior or inferior to those of children reared by their own parents, we could either compare the two means (as Dr. Woolf recommends) or we could carry out an analysis of variance : with either method the answer would of necessity be identical. Where more than two means have to be compared, an exact algebraic identity no longer obtains; but there can be little question that for general purposes the second procedure is at once the simpler and the more efficient. Thus, the contrast between first degree and second degree statistics turns out to be rhetorical rather than real.

For a convincing answer to the more technical details of Dr. Woolf's criticisms, the reader may be referred to Professor Mather's own paper in the same publication ([33], pp. 103–111). As he remarks, there are certain minor but undeniable defects in the particular researches that Dr. Woolf has discussed; and these reduce the cogency of the conclusions reached. But most of the shortcomings seem to arise, not from the inadequacy of the statistical procedure,

but from inadequacies in the data available or in the planning of the experiments by which the data were obtained.

V. Alternative Hypotheses

The Inductive Procedure

In conclusion we should like to draw attention to one point that nearly all our critics have overlooked, namely, the logical form of the argument by which we have defended the hypothesis advanced. As in all empirical problems, the whole argument is essentially inductive, not deductive; and it is convergent rather than serial. Dr. Harrison declares that " the chain of Burt's reasoning can be no stronger than its weakest link : so long as any one of his premisses remains open to doubt, the conclusion that hangs on it is liable to fall to the ground, and the reader is left free to choose some other way of explaining the facts ". But the conclusion was sustained by several chains, not by one. In the paper on ' Evidence of the Concept of Intelligence ', five different proofs were offered—observational, introspective, genetic, neurological, in addition to the mere statistical analysis, which we ourselves regard as no more than confirmatory. Thus the theory we have advanced really rests on what Whewell called ' a consilience of inductions '.

Nor is it sufficient to show that the facts adduced *may* be explained in some other way : the critic must also show that his way is *more probable* than ours. When Dr. Harrison adds that our own explanation is ' never conclusively proved ', we readily agree : in the natural sciences, nothing is ever *conclusively* proved. The final inference was put forward not as conclusive or certain, but merely as more probable than any alternative explanation. The only way to rebut this argument is for the critic to produce a constructive hypothesis of his own, and formally demonstrate its higher probability.

What then are the rival explanations that our opponents have to offer ? We have searched their pages [16], [21], [26], [33]; and nowhere can we find a clearly defined theory, much less any formal proof, which would show, as Fisher for example showed, that the new theory yields a better fit to the facts observed. However, judging from fragmentary comments that have emerged in the course of the discussion, there would seem to be two more or less plausible alternatives that compete with our own. Most of the critics who, like ourselves, accept the possibility that intelligence is in part inherited would apparently favour some form of the venerable theory of ' blended ' inheritance in place of the neo-Mendelian theory of ' multifactor ' inheritance. The rest of our opponents seem to hold that inheritance plays only a negligible or at any rate an unimportant part, and ascribe the observable differences between individuals primarily to environmental influence.

1. *Blended Inheritance.* The theory of ' blended inheritance ' has the sanction of Darwin's name behind it, and was almost universally accepted until the dramatic rediscovery of Mendel's hypothesis of ' particulate inheritance '. The theory of blending is also the explanation which, by its very simplicity,

naturally commends itself to those who are unfamiliar with modern genetics. The facts most frequently adduced in its support are those of ' regression ', which both Galton and Pearson interpreted in terms of blending. Put in explicit shape, the argument runs somewhat as follows. Suppose the general mean of the population to be 100 I.Q. units, and that of the father to be $100+x$ (where x may be either positive or negative); then, on the hypothesis of random mating, the most probable level of the wife's intelligence will be 100; hence that of the child will be the average of the two, namely,

$$\tfrac{1}{2}\{(100+x)+100\} = 100+\tfrac{1}{2}x;$$

thus, as a result of bisexual reproduction, there will be a regression of 0·50. To take account of the effects of assortative mating and remoter ancestry, a slightly more complex version was elaborated by Karl Pearson ([30], p. 456). But the underlying principle remains the same.

Here, however, a mere comparison of means (such as Dr. Woolf recommends) does not enable us to decide between the hypotheses of blending and of multiple factors, since, as we have seen, the value deduced from each of the two theories would (under the conditions stated) be exactly the same. But, when we turn to the second degree statistics, we find a striking difference. For simplicity, let us take the general mean to be zero; and let the deviations of father, mother, and child be x, y, and z respectively, and the correlation between fathers and mothers r. Then (ignoring such complicating influences as dominance and environmental differences) the deviation of the child will, on both hypotheses, be the average of its parents' deviation, viz., $z = \frac{1}{2}(x+y) = \frac{1}{2}(1+r)x$ ([9], p. 115). If there are N males and N females in each generation, the variance of the adults will be $\sum x^2/N$ and $\sum y^2/N = \sigma^2$ say (since with autosomal factors the difference between the variances of the two sexes would be nil). On the multifactor hypothesis, as we have seen, the variances of the children would remain the same ([9], p. 110); but, on the hypothesis of blended inheritance, the variances of the children would be $\sum z^2/N = \frac{1}{4}\sum(x^2+2xy+y^2)/N = \frac{1}{2}\sigma^2$ if mating is random, or $\frac{1}{2}(1+r)\sigma^2$ if mating is assortative. Thus, the effect of blending is to reduce the inheritable variance by roughly one half in each generation. Now the data both for bodily and for mental characteristics make it clear that no such reduction takes place. On this hypothesis, therefore, we are driven to assume, as Darwin did, that some unknown cause or causes must be continually at work to produce fresh ' spontaneous variations '. But all that we know of the rate and nature of mutation puts any such ad hoc assumption completely out of court.

Accordingly, in the absence of more positive evidence in its support, we may conclude that the hypothesis of blended inheritance has a far lower probability than the hypothesis we have proposed.

2. *Environmental Influence.* Those who hold that the mental differences between individuals are the effect chiefly of environmental conditions commonly follow Watson and the behaviourist school, and draw a sharp distinction between

behaviour and structure. " Differences in structure ", we are told, " are, or may be, to a large extent hereditary; differences in behaviour are due almost exclusively to training. . . . Man is, before all others, an educable animal; and conditioning—the simplest form of education—begins so early, even as early as embryonic life, that any differences in hereditary conditions are completely overlaid by the effects of learning ". " Give me a dozen healthy infants ", says Watson, " and my own specified world to bring them up in, and I will guarantee to train any one to become any type of specialist—doctor, lawyer, artist, beggar-man, or thief " ([39], pp. 104f.). These and similar optimistic claims are, we are told by Dr. Fleming and others, largely borne out by the work of Dr. Stoddard and his colleagues at Iowa ([16], pp. 111f.). As a result, " since the opening of the fifth decade of the present century a greater belief in educability has characterized educational thinking ". Nevertheless, Dr. Fleming continues, even so recently as 1931, " it was possible for an official publication to announce " that the psychological evidence " necessitated the provision, not merely of separate classes, but of separate types of schools for pupils of differing capacity "; and she quotes Burt as an " early sponsor " of this view, who has " recently repudiated " it. " It is no longer supposed ", she adds, " that children can thus be classified into a small number of discrete types ". Children who were formerly classified as ' incurably defective ' are now grouped with milder cases of backwardness, and given special training; and even the current doctrine that children who are ' innately bright ' can be picked out at 11 plus for transference to grammar schools is " now thought to be an injustice to those who are not ".

There is a slight misunderstanding here: Burt has never " repudiated " the evidence for separate schools and classes. But for the rest, we should be in fairly close accord with Dr. Fleming's comments on the administrative changes in current educational practice. Yet that is not because we attach less weight than before to the importance of genetic influences. In any case, this " new emphasis on the powerful effects of environment " is by no means so novel as these writers apparently suppose. It constituted the chief explanation for individual differences offered by the ' philosophic radicals ' of the nineteenth century—the founders of the old associationist school which for so long dominated educational theory in this country; and even Darwin at first attributed ' individual as distinct from specific or varietal difference ' to ' environmental pressure rather than inborn endowment ', until, as he handsomely owned, Galton's book on *Hereditary Genius* ' made a convert ' of him[1]. The argument which Galton himself used to controvert ' the excessive reliance on environmental explanations ' could be turned no less effectively against American and British behaviourists of the present day. Since they admit that man's ' unique educability ' is inherited, just like his power of speech, his foetal face, and his lack of a visible tail, is it not extremely probable that innate deviations in the former characteristic (as in all other innate traits) should distinguish different

[1] See the letter quoted in Galton's *Memories of My Life* (1909), p. 290.

members of the race? Why should educability be the sole biological trait which exhibits no individual variation?

How precisely differences in the environment are supposed to act in order to produce the wide differences observable between one individual and another, is hinted at only in the most vague and speculative way. Nor is it altogether clear what exactly its partisans include under the term environment. We are inclined to suggest that there are three main sets of causal process that require to be considered.

(i) *The Acquisition of Knowledge or Skill.* The older British associationist school, and their descendants the American behaviourists, refer almost exclusively to the consequences of what they variously call 'education', 'training', or 'conditioning'. If, they say, the child from a higher socio-economic class generally manifests a higher degree of intelligence, he owes it simply to the fact that he went to a better school or received a better training at home. How then are his superior opportunities supposed to operate? Chiefly, it would seem, though perhaps not solely, by allowing and encouraging him to learn (*a*) certain items of knowledge and (*b*) certain forms of skill. These, it is argued, will fully account for his successes or his failures in the tests that the psychometrist employs.

Certainly with individual tests like those of the Terman–Binet scale, the child from the poorer and less cultured home is penalized in questions that depend on the knowledge of recondite words (as in the Vocabulary Test) or on certain semi-scholastic skills (like reading, writing, and drawing). But the child from the 'privileged classes', brought up, as he so often is, in a sheltered home, is just as likely to be inexperienced in such practical matters as giving change, discriminating weights, or finding a lost ball. Here it is to be noted that, as we have pointed out in earlier studies of such tests, the advantage by no means rests exclusively with the child from a superior socio-economic class ([4], pp. 164f).

As for group tests, it is said, they are usually verbal: they require an ability to read words, to understand words, and sometimes to spell them. A child's performance in such tests, so it is maintained, will consequently depend upon his schooling and the cultural background of his home. But this is a non-sequitur. Because the reaction required by the test includes an activity that has been learnt, it does not follow that success at the test is determined, or even affected, by *differences* in what has been learnt. The activity may be so simple and automatic that, at the age when the test is applied, it gives rise to no individual differences whatever. Any psychologist who is adept in the construction of such tests will scrupulously avoid words or forms of scholastic skill that are likely to be harder for some children than for others; and will so arrange the tasks required that the child's success in each turns solely on his ability to solve diagnostic problems calling for no special training and no prior experience. In a few rare cases, where the child has lacked the ordinary opportunities for picking up even the essential rudiments—as with those who have spent their

lives on barges or canal-boats—the test-results may at times be distorted by environmental handicaps. But, when that happens, the blame should rest rather with the investigator who chose an unsuitable procedure than with the test as such. A school psychologist who is attached to the education authority's staff should have ample opportunity and adequate resources for detecting these exceptional cases; and the need for such vigilance, and the systematic use of various means for checking such possibilities, were points that were fully recognized in the days when backward pupils could be nominated for certification as mentally defective. Hence we feel fully convinced that, in our own data, errors of this kind have been almost entirely obviated by the plan we have adopted.

(ii) *Stimulation.* Many of our critics would be willing to accept these various contentions. But they will explain that what they themselves mean by environmental aids are not so much the items of knowledge or skill acquired at school or elsewhere, but rather the general mental stimulation provided by a cultured home. The ' stimulation ' is supposed to operate partly by giving the child ample scope and encouragement for exercising his latent powers, and partly by " inculcating a higher motivation and habit of accurate and speedy work ". And thus, as Dr. Fleming tells us, " the longer a child remains in unstimulating environments, the lower becomes his performance in tests of what is called intelligence, and the earlier he is transferred to intellectually stimulating situations the higher becomes his measurable response ".

This version of the environmental hypothesis owes its popularity to the widespread belief in ' transfer of training '. Intelligence, it will be seen, is conceived as a kind of faculty, which is strengthened by continued practice and liable to atrophy through disuse. Without wholly endorsing this familiar doctrine in its usual guise, we are inclined to agree that certain tests may in some small measure be influenced by the child's acquired habits of alertness, responsiveness, and the like. At the same time, it is our impression that the effects of such ' stimulation ' are to be traced quite as often among children from the less cultured classes as among those from better educated families. As one of the headmasters told us: " If anything, your tests seem to favour the ' smart alec ' and the future ' spiv ' rather than the child from an intellectual home ". But in any case the effects are slight as well as disputable; and those who seem to make them part of their hypothesis are avowedly relying on mere surmise, and offer no first hand evidence for their views. In the rare instances in which the examining psychologist suspected any such factors, it would be his duty to allow for differences in attitude or background, and amend his assessment in the light of fuller or better information.

(iii) *Pre- and Post-natal Health.* The environmental conditions to which we ourselves should attach the greatest weight are those affecting the development of the child during the earliest year or two of life, including the period spent *in utero*. In certain cases there can be little question that illness, weakness, or malnutrition during these formative stages has permanently impaired not only

the child's physical health but also the growth of his central nervous system. Moreover, the effects that ensue cannot, by their very nature, be eliminated by changing the type of test or by improving the method of assessment. Even among well-to-do families, the first child, and those born when the mother is advanced in years, seem prone to suffer from pre-natal handicaps and perhaps from the inexperience or the incompetence of the mother in her capacity as trainer and nurse. Some indication of the effects that follow may be gleaned by classifying test-results according to the order of the child's birth : with the majority of tests, as one of us has pointed out in earlier papers, the averages for first born and last born are decidedly lower than those for the intervening children. However, with our own assessments such effects appear extremely small (a matter of less than half an I.Q. point). Moreover, during the last few decades there has been a marked decline in the amount of weakness or ill health attributable to causes of this type : that is strikingly shown by the spectacular drop in the figures for infant mortality. Yet, in spite of these improvements, the test performances of the child population have displayed little or no corresponding change in the course of these years; and certainly there is not a shred of evidence to suggest that the proportion of educationally subnormal pupils—dull, backward, and feebleminded—has in any way diminished during this period, but if anything rather the reverse.

But the most cogent argument against these versions of the environmental hypothesis is the existence of children, suffering in a severe degree from one or other of the handicaps described—often from all three—who nevertheless have gained high assessments for intelligence, and even high scores on tests of the usual type. Anyone who may happen to have worked among children in the poorer slums of our larger cities forty years ago will recall numerous instances of this kind. In our earlier surveys we found, among school children classified as living in ' very poor ' or ' poor ' conditions (classes A, B, C, and D as defined in *The Young Delinquent*, p. 68), that as many as 28·4 per cent had I.Q.'s above the general mean (100), 3·1 per cent I.Q.'s over 120, and 0·4 per cent I.Q.'s over 130. All of these came from homes of a deplorably low intellectual status; the majority had received a far more meagre education than the average child—some indeed had made less than half the normal school attendances; and about one-third had suffered from severe malnutrition and ill-health during their earliest years. The obverse of this is equally true. Among children of well-to-do and highly intellectual parents, whose care and treatment have seemed impeccable, much the same proportions have proved to be irremediably dull or hopelessly defective : as evidence, however, such cases are less convincing, since one can never be perfectly certain that some obscure pre-natal disturbance may not have escaped detection.

VI. Conclusions

Our final conclusion, therefore, may be summed up as follows. First, there can be no question that environmental conditions may influence the

results obtained with intelligence tests of the usual kind. How great that influence is will depend on the style of test employed and the type of child examined : with the ordinary school population, and with tests which have been competently constructed, selected, and adminstered, we believe that environmental differences are far less influential than genetic, and can as a rule account for no more than 25 per cent of the actual variance. Secondly, we do not deny that environmental factors of the kinds we have enumerated, particularly the last, may to some extent have affected the observable intelligence of individual children to a degree which no refinements in the methods of testing or assessment could possibly eliminate. If so, however, it still seems clear that such effects must be comparatively slight; otherwise it would not be so difficult to secure direct and indusputable evidence to demonstrate their presence. We therefore conclude that environment can have played only a negligible part in determining the actual figures—the variances, the correlation coefficients, the univariate and bivariate distributions—that formed the main basis of our argument.

So far, therefore, this more detailed examination, both of the objections advanced and of the alternative hypotheses put forward, results in no appreciable modification in the views or the methods originally suggested, but, in the main, tends to confirm rather than to confute the conclusions already drawn.

At the same time it is necessary once again to emphasize that these conclusions are strictly limited and avowedly provisional. The most serious defect in our argument has been the lack of any method for assessing the probable margin of error. To meet the doubts expressed by our various critics, it is necessary not merely to demonstrate the *presence* of genetic factors, but also to determine, if possible in quantitative terms, their practical *importance*. Their mere existence Hogben, for example, is ready to acknowledge; but apparently he would prefer to leave it an open question whether their share in the variance amounts to 70 per cent or perhaps to no more than 10 per cent. That, however, would be useless for practical purposes. With our own data, we believe that a rough indication of the margin of error is furnished by the factor we termed ' unreliability '. But a far more rigorous proof is needed. The next step, therefore, should be to plan a full scale inquiry on systematic lines which will permit the application of such tests of significance as are already available (cf. [28]).

And even if our own estimates are accepted as reasonable approximations, it must be frankly recognized th t the conclusions so reached are valid only in reference to the particular conditions under which they were obtained. They would not hold good (*a*) of other mental traits, (*b*) of different modes of assessment, (*c*) of a population of different genetic composition, or (*d*) of a population at a different cultural level : much less would they hold good if there were any subsequent change (*e*) in the actual distribution of environmental and genetic characteristics, or (*f*) in the influences affecting their mutual interaction.

But, when all is said, mere statistical inquiries are not enough. Nothing like a decisive pronouncement on the issues we have raised can be looked for until more extensive research has been undertaken in the field of psychogenetics.

In our opinion an adequate understanding of the processes involved can be gained only by carefully planned experiments on lowlier animals, where pure strains can be secured and breeding controlled. When we know more about the genetics of intelligence in animals, then we may be able to construct with greater confidence a more exact hypothesis regarding the transmission of intelligence in man. We therefore fully endorse the contentions of a recent reviewer who declares that " it is a matter of shame and regret that only an amateurish beginning has been made by psychologists in using pure lines in fundamental research in the nature–nurture area " ([36], p. 344). Meanwhile, the practical decisions we all of us have to take—educational, social, political, or matrimonial—are bound to have genetic consequences; and, if the findings here recorded have any validity, those consequences may be so far-reaching that the questions involved should no longer be left to hasty improvisation as each arises, with nothing to guide us but popular beliefs or semi-intuitional guesswork, but should be carefully studied beforehand in the light of an explicit and self-consistent theory—the best we can work out at the moment, even though the evidence supporting it may still be incomplete.

References

[1] Broad, C. D. (1947). *The Mind and its Place in Nature.* London : Kegan Paul.

[2] Burt, C. (1912). The inheritance of mental characteristics. *Eugen. Rev.*, IV, 168–200.

[3] Burt, C., and Moore, R. C. (1912). The mental differences between the sexes. *J. Exp. Pedag.*, I, 273–284.

[4] Burt, C. (1921). *Mental and Scholastic Tests.* London : P. S. King.

[5] Burt, C. (1935). *The Subnormal Mind.* London : Oxford University Press.

[6] Burt, C. (1937). *The Backward Child.* London : University of London Press.

[7] Burt, C. (1943). Ability and income. *Brit. J. Educ. Psychol.*, XIII, 83–98.

[8] Burt, C. (1955). Evidence for the concept of intelligence. *Brit. J. Educ. Psychol.*, XXV, 158–177.

[9] Burt, C., and Howard, M. (1956). The multifactorial theory of inheritance and its application to intelligence. *Brit. J. Statist. Psychol.*, IX, 95–131.

[10] Campbell, F. (1956). *Eleven Plus and All That.* London : Watts.

[11] Chesterton, G. K. (1922). *Eugenics and Other Evils.* London : Cassell.

[12] Darlington, C. D., and Mather, K. (1949). *The Elements of Genetics.* London : Allen and Unwin.

[13] David, P. R., and Snyder, L. H. *Genetic Variability and Human Behaviour.* Ch. III in *Social Psychology at the Crossroads.* (Ed. M. Sherif.) Harper.

[14] Dunn, L. C., ed. (1941). *Genetics in the Twentieth Century.* New York : Macmillan.

[15] Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edin.*, LII, 399–433.

[16] Fleming, C. M. (1948). *Adolescence: Its Social Psychology.* London : Kegan Paul.

[17] Glass, D. V. (1954). *Social Mobility in Britain.* London : Longmans and Green.

[18] Haldane, J. B. S. (1938). *Heredity and Politics.* London : Allen and Unwin.

[19] Haldane, J. B. S. (1946). The interaction of nature and nurture. *Ann. Eugen.*, XVI, 37–46.

[20] Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, XXVIII, 476–490.

[21] Heim, A. W. (1954). *The Appraisal of Intelligence.* London : Methuen.

[22] HOGBEN, L. (1931). *Genetic Principles in Medicine and Social Science.* London : Williams and Norgate.
[23] HOGBEN, L. (1933). *Nature and Nurture.* London : Williams and Norgate.
[24] LEWIS, D. G. (1956). The normal distribution of intelligence : a critique. *Brit. J. Psychol.* (In press.)
[25] LUSH, J. L. (1947). Family merit and individual merit as bases for selection. *Amer. Nat.*, LXXXI, 241–261, 362–379.
[26] MADDOX, H. (1956). Nature–nurture balance sheets. *Brit. J. Educ. Psychol.* (In press.)
[27] MARTIN, F. M. (1947). Home background and selection for secondary education. *Eugen. Rev.*, XLVIII, 195–202.
[28] MATHER, K. (1949). *Biometrical Genetics: The Study of Continuous Variations.* London : Methuen.
[29] MILES, T. R. (1957). On defining intelligence. *Brit. J. Psychol.* (In press.)
[30] PEARSON, K. (1900). *Grammar of Science.* London : A. and C. Black.
[31] PEARSON, K. (1906). *Nature and Nurture.* London : Dulau.
[32] PEARSON, K., and LEE, ALICE (1903). On the laws of inheritance in man. I. The inheritance of physical characters. *Biometrika*, II, 357–462.
[33] REEVE, E. C. R., and WADDINGTON, C. H. eds. (1952). *Quantitative Inheritance.* London : H.M. Stationery Office.
[34] SIMON, B. (1953). *Intelligence Testing and the Comprehensive School.* London : Lawrence and Wishart.
[35] SMITH, FAIRFIELD (1937). A discriminant function for plant selection. *Ann. Eugen.*, VII, 240–250.
[36] STONE, C. P. (1947). Methodological resources for the experimental study of innate behaviour as related to environmental factors. *Psychol. Rev.*, LIV, 342–347.
[37] THOMSON, G. H. (1948). *The Factorial Analysis of Human Ability.* London : London University Press.
[38] VERNON, P. E. *et al.* (1950). ap. *The Advancement of Science* (*Brit. Assoc. Ann. Rep.*). London : Murray.
[39] WATSON, J. B. (1931). *Behaviorism.* London : Kegan Paul.
[40] WHETHAM, D., and C. D. (1909). *The Family and the Nation.* London : Murray.
[41] ROYAL COMMISSION ON POPULATION (1941). *Memoranda presented to the Commission.* London : H.M. Stationery Office.

BOOK REVIEWS

The Organization of the Cerebral Cortex. By D. A. SHOLL. Methuen, 1956. Pp. 16+125. 18*s*.

This is a book which should be read not only by the neurologist but by every serious student of psychology. Dr. Sholl's main contention is expressed in his concluding words : " whether the cortex is studied by the anatomist, the physiologist, or the psychologist, the model employed should be based on the concept of probability ". It must therefore be derived from statistical hypotheses and be discussed in statistical language.

He begins with a review of the chief histological and experimental methods available, and illustrates the various microscopic techniques with a series of photographs revealing a high degree of technical skill. On the basis of his own investigations he then criticizes the familiar maps of the cortex which are reproduced in nearly every textbook and are cited as demonstrating structural differences between various regions of the brain corresponding to their supposed differences in function. As Dr. Sholl points out, even the best of such maps were the outcome of subjective criteria and manifestly influenced by the preconceived ideas of the cartographers, particularly their " rather crude psychological theories ". Few recognized " the enormous variability that exists between individual human brains ". Many of the differences, as Bok showed twenty years ago, arise merely from irregularities of growth, differences in local blood supply, and above all the varying amount of deformation produced by the curving surface of the brain.

Dr. Sholl himself holds that the characteristic chiefly calling for study is not the type of nerve cell or the predominance of fibres, but rather " the pattern of connectivity ". This is largely indicated by the manner and the extent of the branching which the axons and dendrites exhibit. These he relates to other aspects of cerebral organization; and is thus led to a most suggestive discussion of the ' quantification of neuronal connectivity '.

To the psychologist the most interesting chapter in the book is that which describes " the mode of operation of the cerebral cortex ". The *ad hoc* explanations put forward by Pavlov, and the somewhat naive theories deduced from the behaviour of ingenious machines like Ashby's homeostat and Grey Walter's mechanical tortoises, are criticized in some detail. Of the two alternative analogies which have dominated most of the older theories—the telephone switchboard with its precise connections, and the field theory treating the cortex as a statistically homogeneous medium, Dr. Sholl inclines, if anything, towards the second : " neural integration must consist of complex patterns of interaction, not of conduction over relatively isolated paths ".

In his discussion of mental abilities like ' memory ' or ' intelligence ', he unhesitatingly rejects the earlier types of hypothesis that tried to allocate them to ' circumscribed regions '. ' Intelligence ', he holds, like ' retentivity ', must be a general attribute of cortical activity, resulting from its over-all mode of organization.

The evidence for these various conclusions is lucidly expounded, and supported by quantitative determinations. However, as he explains, " the statistical mode of description has been adopted, not because of the technical difficulties involved in the study of individual neurons, but because it is the appropriate method to adopt for the study of a system whose interactions can only be described in terms of probability ". CYRIL BURT.

Psychological Tests and Personnel Decisions. By L. J. CRONBACH and G. C. GLESER. University of Illinois Press, 1957. Pp. xii+165. $3.00.

Most of those who have written on the theory of mental tests have assumed that the efficiency of any selection procedure may be judged exclusively in terms of the validity coefficients, regardless of the decisions to be taken. Education committees, however, especially in the early days of mental testing, were apt to inquire whether the *increased* information so gained would justify the additional *cost*. Dr. Cronbach has set out to answer the questions thus raised by developing what Wald has taught us to call ' statistical decision theory '. The book which he and his collaborator have produced contains an admirable account of the construction and use of mental tests from this highly practical standpoint. It should be in the hands of all who work in the fields of educational psychology or personnel selection. W. R. NORTON.